2016

A Study on the Influence of Perceptual Distortion in the Scoring of Musical Performances by Florida Bandmasters Association Adjudicators

Raymond Donato
University of Central Florida

A STUDY ON THE INFLUENCE OF PERCEPTUAL DISTORTION IN THE SCORING OF MUSICAL PERFORMANCES BY FLORIDA BANDMASTERS ASSOCIATION ADJUDICATORS

by

RAYMOND A. DONATO
B.M. Florida Atlantic University, 1996
M.A. Florida Atlantic University, 2000

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Education
in the School of Teaching, Learning and Leadership
in the College of Education and Human Performance
at the University of Central Florida
Orlando, Florida

Spring Term
2016

Major Professor: Kenneth Murray

ABSTRACT

This study explored adjudicator reliability in scores assessed at the Florida Bandmasters Association (FBA) Music Performance Assessment. It investigated how adjudicators under conflicting sets of circumstances interpreted the criteria and rated musical performances. A sample of five concert band audio recordings from the FBA resource library were chosen and a sample of participants were selected to score the recordings using the criteria currently in use by the Florida Bandmasters Association. These participants were chosen from Certified FBA concert band adjudicators, FBA members who are not certified concert band adjudicators and out of state judges who are certified though other judges association. Differences between groups were examined. In addition, data were collected on the participants' ranking of the musical criteria from the FBA concert band assessment instrument.

From analysis of the data, it was reasonable to conclude that there is a significant difference in scoring of musical performances between face-to-face adjudicators who evaluated a live performance, and blind adjudicators who evaluated the same performance via a recorded audio only presentation. This study may provide valuable information that could lead to better development of a fair and balanced rating system.

ACKNOWLEDGMENTS

This has been a lengthy journey to say the least. I owe a great debt of gratitude to my dissertation committee; Dr. Barbara Murray, Dr. Walter Doherty and Dr. Robert Everett for their patience, time and guidance. A special sincere thank you to my committee chair, Dr. Kenneth Murray, for lighting a fire under me and pushing me to finally complete my research when it seemed like all hope might be lost.

Many thanks to Laura Wilcox-Curll and Leah Mitchell in the Graduate Affairs Office for the thankless job of making sure I was on top of countless pieces of paperwork, keeping me apprised of looming deadlines and making sure I was properly enrolled in the right course for the right amount of credits each and every semester.

I am grateful to all the professional educators, musicians and adjudicators I have had the honor of working with throughout my career: past, present and future. Also, many thanks to all of my friends who offered support and encouragement during this extended educational endeavor.

Lastly, but most importantly, thank you my family for unending love and support no matter what foolish, imprudent, irresponsible, hazardous or immature ideas I've come up with in my lifetime.

LGI

TABLE OF CONTENTS

LIST OF FIGURES ix

LIST OF TABLES x

LIST OF ABBREVIATIONS xii

CHAPTER 1: INTRODUCTION 1

Background of Study 1

Statement of the Problem 3

Purpose of the Study 4

Significance of the Study 4

Definition of Terms 5

Conceptual Framework 8

Research Questions 16

Hypothesis 17

Methodology and Data Collection 17

Study Limitations 18

CHAPTER 2: LITERATURE REVIEW 20

History and Justification of Music Education in the United States 20

Florida Bandmasters Association and the Judging Process 24

Developing the Judging Instrument 27

Qualitative Aspects of the Adjudication Process 29

Halo Effect 33

Halo Effect in Teacher Evaluations 34
Director Influence on Musical Performance 36
Measuring Success in Music Education – Director Influences 40
Measuring Success in Music Education – Student Influences 44
Measuring Success in Music Education – Other Influences 44
Measuring Success in Music Education – Teaching Styles 46
Music Education Policy and Law 48
Music Policy Formation & Implementation 50
Teacher Evaluations at the National Level 53
Educator Self Evaluation – Implementing State and National Standards 57
Teacher Assessment In Florida 61
CHAPTER 3: METHODOLOGY 64
Purpose & Background 64
Research Design and Appropriateness 65
Setting 66
Consent Process 66
Participant Process 68
Withdrawal of Participants 69
Risks, Benefits and Participant Privacy 69
Participants and Selection Process 70
Provisions to Maintain the Confidentiality of Data 71
Research Questions 72

Hypothesis 72
Data Analysis 73
Validation of the Survey Instruments 74
CHAPTER 4: RESULTS 76
Introduction 76
Purpose of the Study 77
Population and Samples 78
Test Selection 80
Findings 81
Research Question 1 84
Research Question 2 87
Research Question 3 90
Research Question 4 93
Summary 98
CHAPTER 5: DISCUSSION 102
Introduction 102
Statement of the Problem 102
Summary 103
Research Question 1 104
Research Question 2 106
Research Question 3 107
Research Question 4 108

Conclusions....111

Delimitations and Recommendations for Future Study....116

APPENDIX B: IRB CONSENT FORM....122

APPENDIX C: MUSICAL EXAMPLE EVALUATION FORM....128

APPENDIX D: MUSICAL ELEMENTS ORDER OF IMPORTANCE SURVEY....130

APPENDIX E: FBA ASSESSMENT INSTRUMENT....133

LIST OF REFERENCES....135

LIST OF FIGURES

Figure 1: Scores Assessed by Face-to-Face and Certified FBA Adjudicators 86

Figure 2: Scores Assessed by Face-to-Face and Non-Certified FBA Adjudicators 90

Figure 3: Scores Assessed by Face-to-Face and Certified Non-Local Adjudicators 93

LIST OF TABLES

Table 1: Research Questions, Variables, Data and Analysis 73

Table 2: Pilot Study, Cronbach's Alpha Test Statistic 75

Table 3: Pilot Study, Item-Total Statistics 75

Table 4: Test of Normality For Each Independent Group of Blind Adjudicators[a] 82

Table 5: Kruskal-Wallis Test Statistics For Scores Between Adjudicator Groups[a,b] 83

Table 6: Mean Rank Scores of Each Adjudicator Group 83

Table 7: Mann-Whitney Test Statistics, Face-to-Face and Certified FBA Adjudicators[a] 85

Table 8: Face-to-Face and Certified FBA Adjudicator Mean Ranks 85

Table 9: Face-to-Face and Certified FBA Adjudicator Pairwise Comparison 86

Table 10: Mann-Whitney Test, Face-to-Face and Non-Certified FBA Adjudicators[a] 88

Table 11: Face-to-Face and Non-Certified FBA Adjudicator Mean Ranks 88

Table 12: Face-to-Face and Non-Certified FBA Pairwise Comparison 89

Table 13: Mann-Whitney Test, Face-to-Face and Certified Non-Local Adjudicators[a] 91

Table 14: Face-to-Face and Certified Non-Local Mean Ranks 91

Table 15: Face-to-Face and Certified Non-Local Pairwise Comparison 92

Table 16: Sub-Caption Medians 94

Table 17: Friedman and Kendall's W Test Statistics between Sub-Captions 95

Table 18: Wilcoxon[a] Test Statistics Between Sub-Captions 95

Table 19: Performance Fundamentals Criteria 96

Table 20: Technical Preparation Criteria 97

Table 21: Musical Effect Criteria 97

Table 22: Summary of Findings 100

LIST OF ABBREVIATIONS

CSJA Central States Judges Association

FBA Florida Bandmasters Association

FCAT Florida Comprehensive Assessment Test

IRB International Review Board

MPA Music Performance Assessment

RTTT Race to the Top

VAM Value Added Model

CHAPTER 1: INTRODUCTION

Background of Study

Music has been a standard subject in most public schools since the beginning of compulsory education in America (Tellstrom, 1971). While it is now an established subject, music educators still find themselves defending their programs from curriculum and budget cuts, requiring rationalizing on how music education programs contribute to academic performance across subject areas (Jorgenson, 1995; Myers, 2002). In addition, music education falls outside the area of standardized testing, making it difficult for administrators and stakeholders to properly assess its academic value. While some national standards for music education have been developed, the matching of musical objectives to summative assessment techniques has yet to occur in a uniformed manner (Colwell, 1999). Therefore, music education has found itself with a problem. On one side, there is a general agreement about music education's inherent value to the student, whereas on the other side, decisions regarding such administrative concerns a resource allocation are usually made on the basis of objective, observable and standardized outcomes, and not values or personal bias (Hanna, 2007). Subjects such as math, science, and literature, which are traditionally viewed as more quantitative, may also contain artistic dimensions that are easily overlooked. Many educators recognize that even as these dimensions may not be measured by standardized testing, they are still an important part of their domain (Myers, 2002).

According to Linn (2003), objective assessment of music programs is particularly

difficult because musical outcomes are often judged and interpreted in a subjective manner, with language involving aesthetics, psychomotor skills and performance quality. Reading, writing, math and science, in contrast, are taught and assessed as objective cognitive domains (Linn, 2003). Policy decisions regarding academic programs are usually made on the basis of factual data derived from objective standardized assessment criteria (Porter, 2002). Current forms of music assessment are highly informal in nature, and often subjective, leaving programs without a quantitative method for evaluating their quality. If a music program is thriving, the stakeholders and parents might be happy, and if evaluation festival ratings are good, the program is considered a success (Colwell, 1999). However, beyond those types of informal assessments, the ability to critically and realistically evaluate the quality of various music curricula is severely lacking (Colwell, 1999).

As it stands currently in the state of Florida, ensemble and large group music performance assessments cannot be tied to teacher evaluation, as it is not a measure of individual student achievement. It would take "several years to prove the validity and reliability" of this assessment approach (Florida Bandmaster Association District Meeting #2 Minutes, 2012). As outlined in the new *Student Success Act* as well as *Race to the Top*, the purpose of teacher evaluations are to support student learning through effective instruction. The results can help districts develop school improvement plans and identify needed areas of professional development (Overview of Florida's Teacher Evaluation System, 2016). According to section 1012.34(3)(a)1 of the Florida Statutes, 50% of a teacher's performance evaluation should be based on student learning growth data. The Value Added Model (VAM) currently in use measures these differences in student

performance on state assessments from year to year (Performance Evaluation, 2016). This information is normally collected through mandatory statewide testing, however for teachers of subjects that are not measured by the state, districts assessments can be utilized (Overview of Florida's Teacher Evaluation System, 2016). While districts may choose to use nationally recognized assessments or certification exams, some are choosing to include data from other Florida School Music Association sponsored music events when evaluating teacher effectiveness for the basis of merit pay (Florida Bandmaster Association District Meeting #4 Minutes, 2013). Many music educators believe that the Florida Bandmasters Association should look to align these various assessment tools in a way that, if not directly used for teacher evaluation purposes under the state's current model, might at least allow for "bonus points" to be awarded towards a teacher's effectiveness rating (Florida Bandmaster Association District Meeting #2 Minutes, 2012). However, without a means of assessment that is standardized, music programs will continue to fight for academic legitimacy and scarce resources in an academic environment where accountability is a top priority (Asmus, 1999).

Statement of the Problem

To date, insufficient information exists concerning possible perceptual distortion in scores assessed by adjudicators at the annual Music Performance Assessments (MPA) which school music programs must attend in order to remain members of the Florida Bandmasters Association (FBA). It is agreed generally among school music directors that

the current system provides little feedback in the way of concrete, objective musical criteria and leaves much room for the adjudicator to judge the program based on his or her own biases, as "music is subjective, and there are numerous unique situations throughout the state" (Florida Bandmaster Association Adjudication Committee Report, 2011).

It has been shown through other studies that factors such as director experience, stage presence and choice of repertoire can affect the outcome of a music performance assessment rating. Bias also has been shown in situations where the adjudicator is familiar with the performer(s) or repertoire being performed (Bradley, 1972). In addition, Elliot (1995/1996) concluded that gender stereotypes associated with certain instruments also influenced an evaluator's perception of musical performance in smaller solo or chamber music settings.

Purpose of the Study

The purpose of this study was to determine adjudicator reliability and the degree of perceptual influences in the scoring of musical performances by Florida Bandmasters Association adjudicators.

Significance of the Study

This study examined the criteria contained on the Florida Bandmasters Association concert band music performance assessment instrument, and how an adjudicator under a

contrasting set of circumstances interprets them, which might affect the outcome of a concert band's final assigned rating. The study examined the possibility that subjective factors such as the reputation of a school music program, reputation of a director, band size, age of a director, or gender of a director that are only observed in a face-to-face evaluation can have an impact of the final rating assessed by the adjudicator at a FBA Music Performance Assessment. Furthermore, it is believed that by identifying any inconsistencies, the Florida Bandmasters Association may be better able to properly prepare judges and enhance the learning experience of the music programs that participate, as well as providing a more standardized and objective evaluation method.

There are a few research publications that focus on some observable elements such as the race of the director, ensemble uniform choice, the directors conducting style or even the stage presence of the musicians and how these components can affect the perception of an ensemble's musical performance (Bradley 1972, Elliot 1995/1996). However, there is little to no research that simply tests the reliability of the Music Performance Assessment Ratings Sheets used by the Florida Bandmasters Association during a concert band festival. The results of this study may provide valuable information that could lead to better development of a fair and balanced rating system.

Definition of Terms

FBA: Abbreviation for the Florida Bandmasters Association. This is the governing body for all K-12 instrumental music programs in the state of Florida. Its purpose is to offer public

school music programs promotion and support by providing for director in-service, program evaluation, and student performance opportunities.

MPA: Abbreviation for Music Performance Assessment. This is a non-competitive performance opportunity, hosted by the Florida Bandmasters Association, aimed at providing public school music programs an environment that provides evaluation by trained adjudicators in the field of band performance.

Adjudicator: A trained evaluator in the field of music, appointed by the FBA, whose purpose is to provide a concert band with a rating of its stage performance based on a rubric.

Concert Band: A school music performance ensemble that consists of woodwind, brass, and percussion instruments. A concert band's typical repertoire might include wind band literature, arrangements of orchestral compositions and popular tunes.

Director: The certified teacher of a public school music program that, for the purpose of this study, is responsible for the preparation of a musical performance and will conduct the school's concert band on stage during an evaluation by the Florida Bandmasters Association.

Rating: The final grade given to a concert band by a panel of adjudicators assigned by the Florida Bandmasters Association at a Music Performance Assessment. The rating is assessed based on comparison to a set of musical standards centered on a group's level of experience and musical maturity.

Sub-captions: The three major areas an adjudicator is to consider when evaluating a concert band. On the FBA Concert Band adjudicator sheet, these include "Performance Fundamentals", "Technical Preparation" and "Musical Effect".

Certified FBA Adjudicators: Music judges who are trained and endorsed to adjudicate a concert band performance at a Florida Bandmasters Association Music Performance Assessment.

Non-Certified FBA Adjudicators: Music judges who are not trained or endorsed to adjudicate a concert band performance at a Florida Bandmasters Association Music Performance Assessment, but may be certified in another area. Additionally, for the purpose of this study, this includes members of the Florida Bandmasters Association who are active music directors or educators, but not necessarily judges.

Non-Local Certified Adjudicators: Music judges who are trained and endorsed to adjudicate a musical performance by another formal judges association from outside the state of Florida.

Conceptual Framework

According to Smith and Collins (2009), "People's impressions or mental representations of others are fundamental tools for social life". Such judgments can shape our choice of friends, colleagues, partners, political candidates, job applicants and even family (Smith & Collins, 2009). In an attempt to understand the importance of personal perception, much research has been done to interpret how individuals perceive other people. A good deal of information is known about how these impressions, such as common stereotypes, are used to make decisions about others (Gilbert, 1998). Studies have created a vivid picture of the effect of such impressions, yet many still argue that this understanding does not fully explain the way a person might operate in specific social contexts (Robbins & Aydede, 2008). When there is an interaction with additional groups or individuals, it seems that other psychological processes are utilized as well. Clark (1997) refers to them as "inner representational resources". People might also incorporate second hand information obtained from others instead of simply using firsthand impressions. This new perspective from an individual perceiver might also change the impression of an individual within a group (Clark, 1997).

For a complete understanding of such social patterns and impressions, it is necessary to not only consider a perceiver's firsthand interpretation, but also the larger context of multiple perceivers who are actively sharing information and impressions through social networks and relationships over time (Smith & Collins, 2009). While the term "reputation" may give someone a certain positive impression of an individual, most of

the time someone will have his or her own unique opinion of another person. Therefore, someone's reputation might be influenced on whether people generally agree, or disagree, on their impression of that person. These social perceptions involve the perceiver to be actively involved, and the perceiver may choose how much information he wishes to obtain. However, in actual social situations, many competing factors could influence how knowledge and observations of a person are interpreted, or the choice to use that information at all (Smith & Collins, 2009). Perceivers also must elect to obtain more information about a "social target", and often that choice will usually be based on the impression they already have. Subtle types of social avoidance could limit the amount of relevant and meaningful information one might gather about a subject (Fazio, Eiser, & Shook, 2004). If an initial impression leads you to believe someone is rude, you may never seek out a second interaction with him or her. A mistaken undesirable first impression might never be fixed. This might continue to guide one's decisions about interactions with either that individual or members of a social or professional category. Even a concrete positive firsthand experience might not change anything (Fazio, Eiser, & Shook, 2004).

Denrell's (2005) model holds other implications as well. In his opinion, obtaining information about a person without regard to one's current impression will tend to make impressions more positive. He believes this is because, on the average, impressions are negative so a forced exposure will generally be more positive. In this case any prejudice would be reduced, and as the extent of interactions is increased the more optimistic the impression becomes. The assumption that perceivers might decide whether to seek further information on the basis of their current impressions is believed to hold true for Chaiken

(1987) as well. He believes that perceivers will continue to process or seek new information, until they reach a threshold they are confident of, to make a judgment. Therefore bias due to what he calls "selective sampling" might occur whenever an initial impression influences the probability of continued sampling from the observer. This decision to gain more information about the target is only one step, however. Someone must first choose what information about a person he or she is looking for in order to form that initial impression (Chaiken, 1987). In many cases this amounts to finding information that would otherwise not have come into being at all. Many aspects of the perceiver's choices and decisions, as well as other characteristics of the setting where the interaction takes place, could sway what information is elicited (Chaiken, 1987).

Those who expect a target to act a certain way can often influence behavior that will back up those expectations. A perceiver might have an idea about a target that he or she wishes to test by soliciting certain relevant information. This can be done by asking questions that would give positive answers to their hypothesis (Snyder, Tanke, & Berscheid, 1977). In addition, a perceiver's goal for a particular subject might influence the information that is elicited. If someone expects to interact with another person on a short-term basis, with a specific outcome in mind, he or she might pay special attention to only certain information (Neuberg & Fiske, 1987).

Perceivers, who are varied in height, age, or physical attractiveness, will produce different behaviors from social targets (Reis, Nezlek, & Wheeler, 1980). Likewise, perceivers' ethnicity, occupation, or gender might influence the ways others act toward them. For example, Reis, Senchak, and Solomon (1985) concluded that people's everyday

interactions with women were more intimate and personal than interactions with men. Certain facets of a perceiver's personality might also influence others' behaviors in an interaction between them (Thorne, 1987). In fact, as Buss (1987) has pointed out, many commonly known "personality traits" are actually just terms describing everyday reactions to individuals. Perceivers who could be described with such traits will elicit consistent behaviors from others, in turn influencing the impressions that the perceiver forms. Lastly, interactions in different social settings might also restrict social behavior, leading to the formation of a completely different impression (Malloy, Albright, Kenny, Agatstein & Winquist, 1997).

However, targets also have personality differences that will shape behavioral tendencies in a consistent manner. In traits such as agreeableness and conscientiousness, it has been found that people exhibit a reasonable degree of uniformity in their behavior (Craik, 2008). Kenny et al. (2001) estimated that across various types of one-on-one interactions, there are great consistencies in the way an individual behaves, even with different people in different interactions. Here, different perceivers will agree to an extent on who is more congenial, conscientious, or pleasant.

When a perceiver decides to interact with a subject and gain information, the material must still be interpreted. If multiple perceivers gained exactly the same information from someone, they would still most likely interpret it differently. This is because perceivers view subjects through the "lens of their preexisting knowledge structures" (Gilbert, 1998). Rather than being an unbiased view of the target's characteristics, an impression is usually developed by the receiver. Many studies share the

idea that perceivers with different self-ideals will also differ in their typical opinions of others (Gilbert, 1998). The perceiver's power will also influence the way he or she interprets information about a subject. A position of power can lead to more abstract thinking, while a low-power position might encourage more concrete and detailed approaches (Smith and Trope, 2006). Mohr and Kenny (2006) have examined the way that perceivers use common "person models" (an integrated collection of traits) in making an impression of a subject. The researchers found that once a certain model is adopted by an observer, it is used consistently, and often impacts future information about a subject.

People are connected to each other in a way that keeps information flowing within a group, and allows people to share their impressions with each other. Individuals are linked through social and professional networks, and they become connected to each other through friends, acquaintances, and coworkers (Wasserman & Faust, 1994). These ties give perceivers access to information about people they have never directly met, yet still form impressions about. If two perceivers have the same initial ideas of a third-party, impressions are likely to be analogous. Mason et al. (2007) stated that the structure of a social network of people would influence the speed in which information can reach everyone in that network. Also, the presence of any connections between different groups of perceivers who may know the same subject can influence the extent to which impressions between groups are either similar or distinct. This illustrates that social structures and ties between individuals can be just as important as a one-on-one subject to target process of forming an impression (Malloy et al., 1997). Information that is shared between groups can generally make impressions of a target more similar. Communications

tend to slant towards information about a target that is consistent with the audience's known or assumed attitudes (Higgins & Rholes, 1978). This type of biased communication can help solidify an existing impression of a target, and delivering a biased message might also affect the source's own attitude towards the target. This in turn will more closely align both audiences' attitudes about the subject. Stasser & Titus, (1985) have demonstrated that colleagues of decision making units have a propensity to focus their discussion on items of information that are shared by most of the members of the group. This sharing of information may help the perceivers themselves feel closer to each other. Exchanging such information can lead to individuals feeling that they are closer to each other, and possibly give a feeling of superiority towards the target (Bordia & DiFonzo, 2005). Consistent with this, it has been shown that two perceivers, who are friends rather than professional acquaintances, will have similar impressions of other people they both know (Kenny & Kashy, 1994). Exchange of information about subjects throughout a social network allows the group to gain a consensus on an individual quickly and efficiently. This would not be as fast if the perceiver set out to form impressions on his or her own (Fiedler, 2000). In addition, this exchange of information allows each perceiver to combine larger bits of information and will lead to more accurate and reliable impressions. This would not be as easy if each individual was limited to the small samples of information he or she was able to collect personally (Fiedler, 2000).

As Craik (2008) noted, a social network can operate almost as a monitoring system where an entire network could learn about a subject in a more efficient manner than a single perceiver could. However, sometimes the beneficial effects of collective decision-

making in general are limited by groupthink, and a shared group impression may prematurely bring someone to a conclusion without adequately considering all available information (Mason et al., 2007). This is especially likely if someone fails to use information gained personally and instead focus on only that shared information (Stasser & Titus, 1985). In addition, social courses of knowledge may not be accurate if the information itself is inaccurate or biased. Many of the individual and social functions of a group impression depend on the information being, for the most part, relatively accurate (Craik, 2008). Some studies have shown that "rumors are almost always accurate", however, people can also influence gossip by spreading false or exaggerated material to boost some and criticize others (DiFonzo & Bordia, 2007). This behavior could be magnified when a person considers him or herself better than average at judging someone's character, or obtaining information about a target. They would then rely on their own impressions and generally dismiss others if they disagree (Alicke & Govorun, 2005). The manner in which a perceiver elicits character traits and interprets bias can shape the way he or she forms impressions about others. Because these biases are consistent from the perceiver, it might be difficult to become aware of their existence (Griffin & Ross, 1991). In fact, studies have shown that people tend to not see themselves as influential to a target's behavior, even when the influence might be extremely clear and obvious. They are then unlikely to try and correct these self-induced biases. However, when information is obtained from third-party social or professional sources, it becomes more likely that one could be aware of it, and even attempt to correct it (Gilbert & Jones, 1986). Kenny et al. (1994) found that a perceiver should not give credit to another's impression of a subject by only considering the amount

of information available. That amount usually has little relation to truthfulness. Also, a third party may interact with a subject in a different framework than the perceiver, such as the difference between social and professional settings. Malloy et al. (1997) found people who knew targets from the same social context such as work or family generally agreed, but much less so across contexts. Therefore, if you are attempting to form an impression of a professional colleague, information from someone who only knows them on a social level may not help you correctly form your impression. It will be unhelpful in helping you perceive the subject in the context of a work environment (Smith and Collins, 2009).

One fascinating possibility of note is that of "pluralistic ignorance". While believing that most everyone likes a particular target, one might recognize that he or she personally does not. This pattern might be sustained by social alterations and as people change their discussions about a target to match the attitude they assume the group to hold, it might incorrectly confirm the group's belief that "positive impressions are consensual and therefore that their own personal negative impressions are deviant" (Higgins & Rholes, 1978). Neither gossip nor reputation has had much study in social psychology (Foster, 2004). Considering the importance in defining the social context we find in our lives, this is quite surprising. Reputation might even be a proven difference in the way individuals place emphasis on their own self-perception and judgment, and how they value the reputation of a subject (Heine, 2001).

In general, it seems that it is difficult for perceivers to be aware of and to account for the various causes of bias that can affect their impressions (Wegener & Petty, 1997). This can be true in the simplest situation with a one-on-one perceiver and subject relationship,

and the chance of bias is increased with information that has traveled an unknown string of parties through the social network. Given the difficulty in working with and correcting biases, some might question if we can ever expect shared impressions to be accurate (Fielder, 2000). Even though manipulation in an unintentional way can insert false information into a social or professional network, the available evidence suggests that, in general, reputations are substantially accurate and the gain from the aggregation of this information might just outweigh the source of bias (Fielder, 2000).

Research Questions

1. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind certified FBA adjudicators?

2. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-certified FBA adjudicators?

3. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-local certified adjudicators?

4. How do adjudicators rank the importance of the three major sub-captions and the criteria within each sub-caption?

These research questions were chosen in order to gain a greater understanding of how adjudicators perceive and analyze musical performances. They intend to bridge a gap between previous research into perceptual distortions in musical performances and current Florida Bandmaster Association Adjudication practices. In the current educational climate of teacher evaluations, VAM scores, salary, benefits and job security this may lead to an improved adjudication system that might be utilized in the more untraditional and difficult to evaluate music classroom.

Hypothesis

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: At least one mean score is not statistically equal.

Methodology and Data Collection

A sample of five audio recordings were collected. One was chosen from each of the top five Florida public school concert band directors who have the highest frequency of superior ratings at the Florida Bandmasters Association Music Performance assessment. There are currently 236 concert-band certified Florida Bandmaster Association music judges. These adjudicators are certified to judge in any one of the 21 FBA districts across the state. A sample of 10 adjudicators were selected from this population. Ten non-certified FBA members from the state of Florida, as well as 10 certified adjudicators from outside of

Florida were also be selected. A website link was sent to each of the 30 adjudicators which included for their review:

- An MP3 recording of five separate state-level superior rated concert band performances for their review using the Florida Bandmasters Association Concert Band MPA assessment instrument.
- An online survey corresponding to each of the five recordings, which contain the evaluation criteria to be used. The evaluator was able to evaluate the presentation just as if they were at an actual performance.
- A final survey that asked the adjudicator to rank the sub-captions and the criteria within each sub-caption that are found on the Florida Bandmasters Association Concert Band MPA assessment instrument.

Information such as final ratings, and other musical elements were collected and compared to the information and ratings given by the judges at the initial performance.

Study Limitations

A limitation of this study was the inability to also include performances that did not receive superior ratings at the FBA concert music festival, as these recordings are not readily accessible through the Florida Bandmasters Association recording library.

The medium through which this analysis will be conducted may prove to be in part a limitation of this study as well. As this study was conducted through electronic mail and digital audio formats, it did not allow for the judge to make any assumptions or evaluations

based on aesthetic aspects of the performance, director or musicians on stage such as dress, professionalism or carriage. Study participants would need to either view a video or attend a live performance in order for such items to be considered in any analysis or conclusions. The quality of the digital audio recording used versus a live performance may have also influenced an adjudicator's interpretation. Subtle musical nuances present in a live performance may not exist on a digital recording. In addition, this study was bound by the current regulations, procedures and assessment instruments that are currently in place by the Florida Bandmasters Association.

Lastly, the anonymous nature this study might have permitted adjudicators to be more critical of the musical performance, as opposed to a face-to-face evaluation where the judges are directly held accountable for their scores and commentary to the performers and director.

CHAPTER 2: LITERATURE REVIEW

History and Justification of Music Education in the United States

In 1830, William Woodbridge proposed a basis for music education in American schooling. His essay, "On Vocal Music as a Branch of Common Education", gave rationale that was basically unchallenged for over 150 years (Jorgensen, 1994). These sentiments are echoed by James Mursell's "Human Values in Music Education" in 1934, and even discussed in a 1991 report by the National Commission on Music Education (Jorgensen, 1994). Woodbridge (1831) spoke of "the creator" that provided mankind with the gift of music as a way to praise God. If mankind failed to develop and spread this gift, it would show disrespect and ungratefulness (Woodbridge, 1831). According to Woodbridge, there was a direct connection between what the ear hears and the heart feels. He felt this connection between emotion and music could influence the good and the evil in a person, as he believed every feeling could be expressed by a tone, and every tone stimulates the feeling from which it developed. In addition to those influences, he also believed it refreshed the mind, enhanced a person physically, lifted one's moral character and most importantly, improved academic discipline by enforcing order, union and obedience (Jorgensen, 1994). Woodbridge's ideas on music and social discipline spoke to the needs of the industrial enterprises of the era. In addition to the need for workers, there was also a desire for those with a sense of order and discipline in the workplace. It had been clearly demonstrated in the past that one works more effectively when informed, methodical and orderly (Jorgensen, 1994).

The public also viewed the inclusion of music in schools as an advancement of the church, as congregations used music in their ministries and already understood its importance in the development of childhood and adult life (Mursell, 1934). Woodbridge drew his connections from the works of Plato, Martin Luther and Benjamin Rush (Jorgensen, 1994). Plato had an understanding of philosophy and moral development and felt the arts held the key to understanding, imagination and cognition. Martin Luther made similar arguments and believed in a relationship between music and spiritual development. Lastly, Benjamin Rush made the claim that the study of vocal music in particular would help defend against physical ailments such as pulmonary disease and tuberculosis. These arguments provided a strong justification for the inclusion of music education in the public school system and connected music itself to strong economic and political ideals (LeCroy, 1998).

Later, James Mursell (1934) would echo the ideas of Woodbridge. He felt that the goal of musical study was to enable people to live "stronger, more satisfying, more worthy lives". He believed something taught in public school only held value if it released either a human or spiritual quality. As Mursell stated, a person was not defined by the list of skills they had, or the things they knew, but rather in terms of the type of life they should live. An important part of this was a person's spiritual well-being, and the study of music contributed to this. In addition, group music settings gave opportunities for students to create, perform and relate to one another. If the outcome of education was based on morals, then as Mursell believed, every musical activity was a moral undertaking.

Susanne Langer's publication "*Feeling and Form: A Theory of Art Developed from*

Philosophy in a New Key" (1982), discussed the intrinsic and aesthetic value of public school music programs. She believed there was a large part of human history and evolution that existed outside of scientific dialog and study, and must be expressed through means that are based in human emotion and feeling such as music and art. Leonard Meyers (1956) thought that musical meaning was found within the music itself and the way it elicited emotions. He continued further to say that music was tied to cultural norms, and might hold a different meaning based on the listener's knowledge of that culture, rather than just being a universal experience. He believed music educators should focus on a global view of music, rather than strictly a Western classical perspective. Abraham Schwardron (1967) added that music must also be studied by using correct music terminology; such as one would when studying science, language, math or literature. Music educators felt that rather than simply listening to music, the true primary focus of music education was the study of composition and performance. Elliot Eisner gave attention to music's part in an individual's understanding of other pertinent educational topics, such as social studies and politics (Jorgenson, 1994).

While describing music through the notion of aesthetics appealed to music educators, problems arose when attempting to convince educational policymakers of the importance of music education in terms they could recognize as clear-cut benefits (Smith, 1987). Growing economic troubles in the American public school system intensified this. Both politicians and the public were more concerned with balancing the budget rather than the study of subjects rooted in aesthetics, which were considered to have little practical value (Jorgenson, 1994). Those who were in the best position to speak to the public and

politicians on the benefits of music in the public schools were slowly starting to disappear from the school system. Music supervisors, arts supervisors and even consultants were being removed from the system, and this left music educators without the unity and sense of leadership they were used to (Mursell, 1934). In his essay "Music and the Liberal Education", Peter Kivy (1991) gave the impression that while music had both intrinsic and extrinsic values and was enriching to society, one still could not justify the idea that the study of music in public school is essential. He continued his argument by comprising a list of other subjects that might benefit both personal and corporate growth, yet are not included in public schools. If a subject was to be part of the curriculum it must be essential, and music education lacked any argument that would convince the politicians of its importance (Jorgenson, 1994). With no clear political justification for the need of music in schools, the focus becomes convincing the public that it is essential enough to be included in the school curriculum. While professional music educators are convinced of music's inherent value, they must show its functional side to society and the educational system (Phoenix, 1964). These dual sets of values often do not mesh with ease. This juxtaposition is embodied in a report by the National Commission on Music Education titled "Growing Up Complete: The Imperative for Music Education" (1991). The claim was made by the commission that those who study music will also do well in other academic subjects. They showed statistical relationships between reading, math, spelling, mental abilities, critical thinking, problem solving and motor proficiency to suggest that being involved in music encouraged self-esteem, self-expression, creativity, and self-discipline. The commission's view was that the evidence showed overwhelming extrinsic value in music education.

In the present culture where educational choice is prevalent, it is important for music educators to find ways to strengthen their network throughout the country (Eisner, 1985). The goal of music education should be to enrich the American culture, and to shape the public's understanding to such an extent that the policymakers are pressed to provide environments where music education will thrive. A strong philosophy and a clear articulation of music's place in education will become a powerful argument for its place in the schools (Jorgenson, 1994). Herbert Read (1958) believed that the role of arts education was to enrich both personal and collective experiences and to prepare citizens to "take their place in democracy". Maxine Green in "The Dialectic of Freedom" (1988) made the statement that American education was partially about finding personal freedom, and is a place to find one's own realities, values and self. Green states that the study of the arts is an important part of a person's spiritual and imaginative self, and therefore would impact one's political ideals. This in turn provides the beginnings of future desires for public music education (Eisner, 1985). According to Reed (1958), if music is to have an essential place in public school, a philosophy of music education must be ingrained into one's idea of freedom, democracy, and social value. How such ideas will be shaped remains to be seen, but they foreshadow visions of music education that are "compelling in the present world" (Reed, 1958).

Florida Bandmasters Association and the Judging Process

As federal, state, and local laws move towards stricter standards of accountability in

public schools, assessment becomes increasingly important in the educational process. While school music directors regularly make both formal and informal assessments of the performers and ensembles that comprise their program, they will also take opportunities to have their ensembles judged by outside sources (About FSMA, 2014). In the state of Florida, The Florida School Music Association (FSMA) oversees several different chapters of music association. Before the FSMA was formed in 1997, public school music programs fell under the supervision of the Florida High School Activities Association (FHSAA) (About FSMA, 2014). When the FHSAA was limited to only athletic organizations by the Florida legislature, public school music directors in the state formed the FSMA to supervise the Music Performance Assessment and act in the interest of music educators and their programs (Frequently Asked Questions About FSMA, 2014). As a paid member of the Florida School Music Association, a school music program has the opportunity to participate is music assessments that are hosted and sanctioned by FSMA (About FSMA, 2014).

The Florida Music Educators Association (FMEA) is the state level association for professional music educators. The FMEA has several individual components to meet the needs of the varying types of music programs such as instrumental, choral, orchestral and general elementary music (Frequently Asked Questions About FSMA, 2014). The Florida Bandmasters Association (FBA) is the component for instrumental music directors, and governs the performance assessment of concert band programs (Philosophy and Purpose of the FBA, 2014). The FBA is divided into 21 districts throughout the state. Each district sponsors and oversees several Music Performance Assessments throughout the school year

for marching, jazz and concert bands (FBA Handbook 2014-2015, 2014). At a Florida Bandmasters Association MPA, a concert band performs on stage for a panel of judges who listen to and evaluate each performance. They then present a rating to the ensemble based on a rubric (referred to as the "sheet") that is both developed by an FBA Adjudication Committee and then approved by the members of the association. A final rating is determined by averaging the individual judge's ratings of the performance (FBA Handbook 2014-2015, 2014).

Judge panel size varies depending on the type of ensemble being evaluated. With a marching band, there are at least four judges rating individual aspects of the band's performance such as music, marching, and the overall effect of the program. Two optional additional judges may be used to adjudicate the band's color guard (such as flags, rifles, dancers and majorettes) as well as the percussion section (FBA Handbook 2014-2015, 2014). In the case of a jazz band assessment, a panel of three judges are used who each evaluate all aspects of the musical presentation. This is the same in a concert band setting, with the exception of a fourth judge who evaluates the band's ability to sight read a piece of music. This assessment takes place in a private room once the band has completed its stage performance (FBA Handbook 2014-2015, 2014). A high school concert band that receives a superior final rating has the option to perform again at the state level. This state evaluation is made under more stringent standards than the district performance events (FBA Handbook 2014-2015, 2014). There are four objectives outlined by the Florida School Music Association. These include "realistic and constructive" evaluations of both solo student performers and large ensemble performances (About FSMA, 2014). In addition, the

Florida Bandmasters Association has also created guidelines in which an FBA member can become a certificated judge. As outlined in the FBA handbook, after having 7 years of teaching experience and after receiving straight superior ratings three out of the last five years, a music director may apply to become certified (FBA Adjudication Handbook 2014-2015, 2014). To begin the process of certification, a director who meets the above requirements must first be nominated by their FBA district members. An application process follows which includes obtaining three letters of recommendation from other current FBA judges. The completed application and letters are reviewed by the FBA Adjudication Committee and then approved by the Executive Board. At this point, the internship process begins, where candidates attend official training and shadow other certified judges during a number of FBA sponsored Music Performance Assessments. At these events, the intern will compare their assessment and ratings with those of the certified judges on the panel and be reviewed by the certified judges. At the culmination of this process (which generally takes about a year), the candidate's materials are sent to the FBA Executive Board for approval, and they will be added to the list of official judges used by FBA for district events (FBA Adjudication Handbook 2014-2015, 2014).

Developing the Judging Instrument

Fiske (1983) describes instrumental performances as aural events that move through time. He believes it is a difficult challenge for adjudicators to listen to a musical performance and be specific about what they have heard. Both music educators and

adjudicators make an attempt to observe many separate levels of musical and technical ability during a student or band performance (Saunders & Holahan, 1997). In a classroom setting, teachers make decisions about aspects of the performer's musical contributions and provide feedback and instructions for improvement. Making judgments, interacting with students and making musical decisions are all basic components of being a music director (Fiske, 1983). Directors try to be unbiased when selecting students for an ensemble, keeping focus on musical attributes (Burnsed, Hinkle & King, 1985). In the case of solo performance evaluation, as in a concert festival, judges are asked to use a rating instrument, known as a sheet, to assess a student's musical ability. Usually, a typical rating sheet has adjudicators rate a musical performance based solely on their own personal ideas of quality (Saunders & Holahan, 1997). An adjudicator is expected to assign a final rating, usually a number or a letter, as an indicator of his or her perception of a musical performance. While overall ratings have been found to be reliable between judges, a specific criterion of rating has not. In addition, rating instruments that use a typical ordinal scale do little to indicate specific qualities and characteristics of a performance that lead an adjudicator to make a decision (Burnsed, Hinkle & King, 1985). Little diagnostic feedback is given with respect to specific performance standards, and directors and musicians alike have little indication of what makes their performance great, fair or substandard (Saunders & Holahan, 1997). Jones (1986) and Winter (1993) have developed judges' sheets that include a Likert scale to gauge an adjudicator's level of agreement towards particular performance criteria. Judges, in this format, were asked to specify along a 1-5 scale the amount they agree or disagree with statements that describe an aspect of a musical

performance. Knowing to what degree an adjudicator's opinions coincide with specific criteria show what the judge thinks about the ensemble's performance abilities (Azzara, 1993).

Rating instruments that were more criteria specific and provided increased feedback from the judge have been developed (Saunders & Holahan, 1997). These types of evaluation sheets include written descriptors of performance levels in which adjudicators describe what they are hearing during a performance without stating if they agree or disagree with how the performance meets typical musical standards (Azzara, 1993).

Qualitative Aspects of the Adjudication Process

As the stakes increase in the area of teacher evaluation, accountability and testing in public education, it has become important that the assessment tools for music educators be handled in a fair manner. The assessment of a director's music ensemble can hold a role in the evaluation by an administrator, a director's job security, recruitment and retention for his program and an overall sense of job satisfaction. For this reason, it is important for the judges to be aware of all the contributing factors when assessing a rating at any Music Performance Assessment event. There is an attempt to "balance and synthesize" qualitative aspects of musical performance in the attempt to provide some kind of grade, judgment or rank (McPherson & Thompson, 1998). They also note four factors that will influence a musical assessment. The first of these is the type of performance judged such as a rehearsed piece of music, something that is improvised, or music read for the first time on

sight. The size of the performance, whether it is a solo performer, or a larger ensemble, is the second factor. The environment is the third factor, and can include a music room setting, a stage presentation or even a one-on-one musical showcase. Finally, the purpose of the assessment comes into consideration, such as an audition, festival or competition (McPherson & Thompson, 1998).

Most concert bands in America participate in music evaluation festivals where they are judged on their prepared stage performance. Usually a set of three adjudicators will independently award an overall rating to an ensemble, and often give a few written comments as well (Bergee, 1995). Music directors place a large importance on these music festivals, and in the age of accountability, a director's future may be based on the outcome. Because of this, the subjective nature of these types of evaluations had always been a concern (Burnsed, Hinkle & King, 1985). A study by Fiske (1983) found faults in inter-judge reliability and recommended a panel of at least seven judges to help establish consistency between judges. Burnsed, Hinkle & King (1985) found that judges disagreed significantly in certain captions, with tone quality being the most notable.

It has been shown through various other studies that factors of a "non-musical nature" can come into consideration when evaluation a musical performance. VanWeelden (2002) ran a series of studies investigating such criteria as a conductor's build, race and gender to establish if there was a correlation between these factors and the ratings assessed by music performance judges. The research showed that female directors with a thin build were higher rated in musical performance than those with a larger build (VanWeelden, 2002). When conducting a traditional African American Spiritual, it was

shown that the race of the conductor was also a significant factor. Concert bands conducted by African American conductors were rated higher than ensembles led by white conductors even though the musical performances provided to the judges were identical (VanWeelden, 2004). Further research by VanWeelden continued to investigate the effect of racial stereotyping on conductors and their music. Two musical excerpts were provided, a performance of typical western concert music and one from the same African American Spiritual in the 2004 study. The judges rated white conductors higher with respect to the western concert band literature, and African American conductors higher when leading the African American Spiritual, leading the researchers to the conclusion that the judges may have racially stereotyped the conductors (2004). Further study by VanWeelden in 2007 found that race of the judge did not play a role in assessment. Both white and black judges were consistent in their assessment of conductors of Western vs. Spiritual music.

Other studies have examined the role of the evaluator in Music Performance Assessments. Bradley (1972) found that a factor such as a judge's personality, training, experience, knowledge of repertoire and familiarity with the musical performer all strongly affect the outcome of the adjudicator's musical assessment. Elliott (1995/1996) however, discovered that typical gender stereotyping of instruments often influence a judge's perception of a musical performance, without regard to the judge's training and experience in music. However this was only the case with female performers, as male performers scores did not change based on the perceived gender association of their instrument (Elliott, 1995/1996). Once again, in the case of these studies, the musical performance given to the judges were the same throughout, heightening the fact that gender was a

consideration in the musical evaluations.

According to Bergee (1995) the problem might lie with the vague criterion that is given for the purpose of evaluation. The judging process involves human ideas and perceptions of music characteristics, which leaves much room for interpretation (O'Brien, 1992). Nunnally (1978) believes that defining a valid way to measure means focusing on the words "Concert Band Performance". Some researchers, such as Burnsed, Hinkle and King (1985) and Fiske (1975) feel that only large criteria such as overall musical effect would be sufficient to rate a performance, while others such as Abeles (1973) and Bergee (1993) feel that music is complex, and the measurement tool must be equally as complex. A successful musical performance is a united, cohesive phenomenon and detailed feedback is needed to properly assess it (Fiske, 1975). However, a more general type of assessment sheet is usually used, and does little in the way of providing feedback to the director. Judges are usually encouraged to add additional written comments to the adjudication sheet (Bergee, 1993). This type of detailed criteria feedback is considered to be the essence of festival music adjudication, yet most directors only focus on the final categorical rating (Neilson, 1973). Neilson adds that this method of adjudication relies on assumptions such as stability across time and performances, captions will not be taken out of context, few captions are enough to adequately judge a performance, and broad concepts such as technique, tone, etc. are universally understood. While Burnstead, Hinkle and King (1985) found a correlation between a band's final rating and the captions used on the adjudication instrument, Wagner (1991) did not. He believed that broad captions were lower level ways

to assess the music program, while the specific terms and musical definitions were the most effective.

Halo Effect

An evaluator's tendency to overemphasize the relationship between a subject's traits or behaviors has been called logical error, illusionary halo, correlational bias, and most notably, the halo effect (Feeley, 2002). These various labels notwithstanding, errors thought to be made in an evaluation are still the same. An evaluator fails to differentiate between independent and separate aspects of a subject's behavior or traits (Saal, Downey & Lahey, 1980). The halo effect is believed by many researchers to exist in most data sets that involve ratings of people by other people (Feeley, 2002). According to Feldman (1986), halo errors seem to be inevitable. Others such as Kozolwski (1986), Pike (1999) and Cooper (1981) have dubbed it both global and consistent. The largest problem with halo effect is the amount of weight that is given to, or subtracted from, subject's scores based on these evaluator's perceptions (Feldman, 1986). These errors tend to lower the validity of a subject's rating, and the real world decisions such as performance evaluations, employee selection and teacher evaluations could be affected by these halo errors (Feeley, 2002).

In a general sense, halo errors stem from an evaluator's general impression of a subject, and basing their evaluation of other independent attributes on that impression (Feeley, 2002). Halo is usually thought to be a direct function of the "cognitive processes of the rater" (Murphy & Anhalt, 1992). According to Fisicaro and Lance (1990), there are

three models of halo error. The general impression model states that an evaluator's general idea of a subject can influence his or her judgment in other independent unrelated dimensions. The salient model suggests that assessment of a person in one area can influence an assessment of that same person in another area, even if the variables are unrelated. Finally, the inadequate discrimination model is defined as the evaluator's inability to distinguish between clearly distinct aspects of a subject (Fisicaro & Lance, 1990). The research of Cooper (1981) explains several other bases of halo error. An evaluator who is unskilled or only samples a small part of the subject tends to make more halo errors. With little information to use, evaluators use more of a global impression, which might link irrelevant information to the area being rated. In addition, Cooper found that an "abstract and fuzzy category" might result in a larger halo effect than a category that is clearly defined. This would force the evaluator to "lump together" any groupings that lack clarity (Feeley, 2002). Lastly, Cooper (1981) believes that halo effect is directly related to any lack of effort or carelessness displayed by the evaluator. This apathy forces the evaluator to extend known information across multiple elements of the subject (Feeley, 2002).

Halo Effect in Teacher Evaluations

The study of teacher evaluation has a lengthy history (Darby, 2007). As early as 1974, Bassin used a set of Likert scales to examine five aspects of teaching as evaluated by college level students. He found that courses dealing with quantitative matter receive lower

overall ratings of than those of a qualitative nature. Pohlmann, in 1975, while looking at teacher and course characteristics, concluded that students enjoyed elective classes more than required ones. Research into teacher and course evaluation continued with Rae (1997) and Shevlin (2000). While the assumption is that within these evaluations the scales are independent of each other, the value of the scores might be reduced if, in fact, sets of responses are influenced by reactions from another set (Darby, 2007). Research by Cohen (1981) searched for correlations between specific, predicted areas where he felt there would be a connection, such as a student's impression of an instructor and the grade the student received in the class. However, more recent studies on teacher and course evaluation have overlooked this issue of the independence of scales. While the concept of the halo effect is known in the field of perception, it is not an idea that is usually applied to teacher evaluations (Darby, 2007). According to Blum and Naylor (1968) the halo effect is described as the inclination to allow one trait of an individual have an influence over other traits of that person, having either a positive or negative influence. A problem in showing if the halo effect has occurred can be that various traits are related and based on actual similarities rather than just a social influence (Thorndike, 1920). Researchers Mi-Young and Jyotika (2003) believed that even as the halo effect exists, proper steps could be taken to minimize its effect. Such steps might include significant differences in the items evaluated and, according to Kobrynowicz and Biernat (1997), incorporating more open-ended response sections. They believed these free responses allowed for more expression than a typical scaled response. It seems as individual Likert-type evaluation scales are not seen as independent factors by evaluators, and a halo effect occurs (Darby, 2007). In

addition, the lack of connection between scaled and open-ended responses suggest that evaluators react to each of these formats quite differently (Darby, 2007).

In a study by Asch (1946), it was demonstrated that adding descriptive qualities to a hypothetical person such as "warm" or "cold" could modify an evaluator's perception of that person. In addition, he found that other qualities, such as describing a person as "polite" did not alter the evaluator's impression of the hypothetical person (Widmeyer, 1988). Kelly (1950) continued the work of Asch by demonstrating that the same "warm" and "cold" descriptors could influence an evaluator's perception of a real person whom they actually had an interaction with. In this study, Kelly found that evaluators who were told that a subject was "warm" gave better ratings of the subject's personal qualities than the evaluators who were told that the subject was "cold". Lastly, Kelly also discovered that evaluators were more likely to participate in discussions with the subject if they considered them a "warm" person. This work of Kelly (1950) and Asch (1946) inspired a large amount of research in perception based on various characteristics of both the evaluator and the subject (Widmeyer, 1988).

Director Influence on Musical Performance

A concert band's performance is the result of not only the musicians on stage, but also their reaction to the conductor (Morrison, Price, Geiger & Cornacchio, 2009). Add not only the audience members, but also the perspective of the evaluator, and the combination of "actions, sounds, and the larger context in which this all takes place" (pg. 37) becomes

even more complex (Small, 1998). In the end, an evaluation of a live performance may only be somewhat attributed to what is heard when you include variables such as conductor cues and actions (Morrison et al., 2009), and one might judge an ensemble's musicality based on the expressiveness of the onstage director. The most obvious of these musical and visual associations are shown in the relationship between tempo and rhythm, but it is believed that motion can also affect melody and harmony (Shove & Repp, 1995). With respect to smaller chamber and solo performances, Davidson (1993) stated that there is little literature discussing the visual contribution of the conductor to the performance, and he believed that this lack of evidence also applied to large stage ensembles. Clark (2005) added that we associate sounds with motion and these interactions have not received much attention in the study of music.

More recently, Vines, Krumhansl, Wanderley, and Levitin, (2006) have studied the connection between musical tension and emotion with that of the performer's movements onstage. They found a relationship between the visual movement of the performer and the music phrasing that is perceived. In a study by Juchniewicz (2008), it was found that evaluators gave higher ratings on musical criteria such as dynamics, rubato and phrasing to those performers who incorporated full movement of their body to their performance in comparison to similar musical performances that contained little to no body movement. It was also reported by Thompson, Graham, and Russo (2005) that performers could communicate expressiveness through facial expressions, in turn enhancing an evaluator's listening experience. They also added that with facial expressions intentionally expressing positive or negative emotion, listeners rated the music happier or sadder accordingly. As

the art of conducting is usually in part to convey musical characteristics through visual movement, one might expect that a live performance utilizing a conductor would provide more information to the listener (Morrison et al., 2009). Geringer, Cassidy, and Byo (1997) found that study participants reported higher scores on a listening test when watching a live performance, rather than those who watched a video of the same musical performance set to animation.

Motions of the musical conductor, according to Bram and Braem (2001), are types of "visual metaphors" that can differ depending on the performance situation. Gestures given in a performance may vary from those given during a rehearsal (Garnett, 2005). There does not seem to be a well-established relationship between the gestures of a conductor and the resulting performance (Morrison et al., 2009). Byo (1990) found that some gestures are usually tied to specific musical ideas, and that experienced conductors both understand and utilize this. Still, any real relationship is unclear (Morrison et al., 2009). It has been found that a musician's performance is more accurate, yet not as expressive, when watching a conductor on videotape (Sidoti, 1990). On the other hand, in a study by House (2000), it was found that individual performances by more advanced players became more expressive as a videotaped conductor became more expressive. Music students in the 8th grade were found to have no real measurable difference in their performance regardless of the expressiveness of the conductor's gestures (Price & Winter, 1991). This might lead to the idea that as performers become more experienced, they become more sensitive to the expressive details of a musical conductor (Morrison e al., 2009).

Those who found more of a correlation between perceived conductor effectiveness and ensemble musicality include Van Weelden (2002), Grechesky (1985) and Liab (1993). Liab concluded that audio recordings of performances by expressive conductors were preferred over those with non-expressive conductors by a panel of independent adjudicators, even though they could not see the conductor. It was also discovered by Liab that members of those ensembles had more optimistic opinions about the conductor if he used expressive gestures. Price and Chang (2001, 2005) in a series of studies found that judges who were asked to evaluate both an ensemble's performance as well as its conductor, commented mostly on the expressiveness of the ensemble and conductor, even though they were asked specifically about the quality of the performance. Studies have demonstrated that an evaluator's perception of a musical performance can be affected by conductor attractiveness, gender, race (Elliot, 1995/1996) stage presence and even attire (Wapnick, Mazza, & Darrow, 1998, 2000). Good conductors reflect their interpretation of a piece of musical literature through their presence and movement on stage. However, what is not usually measured is how much the appearance of the conductor actually affects the performance of the musical ensemble (Morrison et al., 2009). A more direct question might even be whether the movement, appearance or presence of the conductor has a direct influence on what an evaluator or audience member believes they hear. The communications of the onstage director are not only directed towards the musician, but to the listener as well (Elliot, 1995/1996). Morrison et al. (2009) believed that the energy and movement of a passionate conductor could enhance the qualities of the music being perceived by an evaluator or audience member.

Many aspects of an ensemble's verbal behavior, nonverbal behavior, attitude and even conductor approach have been examined in relationship to a band's performance and opinion of conductor ability (Fredrickson, Johnson & Robinson, 1998). Sheldon (2000) even believed that the conductor's overall disposition could influence an evaluator's perception of the music. It is generally agreed that the expressiveness of a music ensemble holds the most weight with respect to a band's final rating at a concert evaluation (Burnsed & King, 1987). The quality of the selected music was also a contributing factor. Lucas, Hamann and Teachout (1996) studied the effect of presentation modes (audio only, video only and audio/visual combined) on an evaluator's feeling on performance quality and expressiveness, finding that the largest difference was between video only and audio only modes. Wapnick, Darrow, Kovacs, & Dalrymple (1997) expanded that area of study and found that the attractiveness of the conductor played a role in the evaluation of a prerecorded solo performance. The results found that women scored higher than men and an attractive performer was given a higher rating than an unattractive one. According to Price and Chang (2005), research in this area could very well be relevant to the evaluation of conductors.

Measuring Success in Music Education – Director Influences

There are many ways to measure student success in the music classroom. A school's music director might consider a highly musical and professional performance as one such way, especially if a panel of certified music adjudicators view it as such (Burnsed, Hinkle &

King, 1985). This might provide a director with an intrinsic reward, such as pride in his program, and other external rewards such as recognition, admiration by his or her peers and in many cases, job security (Beaver, 1973). There are many ways that a musical performance could be considered successful; audience reaction, reaction from the performers, positive comments and reviews and high ratings from at any type of evaluated performance (Dawes, 1989). In the attempt to make their music programs the best they can, and achieve the highest level of musical excellence, most music directors will attempt to understand how their program relates to others programs that have already found success (Davis, 2000). According to Goodstein (1984 & 1987) variables such as music program size, funding, administrative support and experience of the director can all factor into the success of a music program. While any of these factors can influence the success or failure of a program, Groulx (2009) states that it is usually more dependent on director factors. There are, however, gaps in the research literature in areas such as teaching style and director personality and how it relates to the success of the music program. These factors are not easily observed, and are difficult to measure in relation to more quantifiable elements like enrollment totals, years of experience or student retention (Groulx 2009 & 2010). The research may not always be concerned with qualitative aspects of a director, as public opinion shows that director personality and ethical values are critical to a program's success and that a teacher's personality can have a large effect on ability to succeed professionally and motivate students (Colwell, 2006).

According to McCrae & Costa (2003) fundamental personality traits such as attitude and beliefs are unlikely to change through the time it takes to earn a music education

degree and during adulthood. However, having an awareness of these traits can help a director overcome any negative effects they might have on professional performance. Even though a music director teaches in the same manner that he or she is accustomed to and were taught themselves, it is possible for him or her to reflect and gain a better understanding of his or her weaknesses and how to improve them (Fontana, 1977 & 1986). Student performance, according to Gumm (2003), can be directly affected by a director's teaching style and he finds it important to uncover any connection between a music program's achievement and that teaching style. Austin (1988) believes that while performances done for a panel of adjudicators can be a foundation for a prideful music program, too much concern with such competitions can be a detriment to broader musical goals. A director who focuses much of the school year perfecting a few pieces of music in preparation for a concert evaluation might produce a technical performance, but it would lack expressive and musical aspects (Croft, 1984). In turn, this might limit the students' introduction to a wider range of musical literature or concepts and decrease their ability to sight-read (Harris, 1991). There is literature to support that a music director's teaching style might not only have an impact on the way he or she rehearses and prepares, but also will ultimately affect student achievement and ratings at adjudicated music festivals (Costello, 2005; Davis 1998; Yarbrough, 1998). Three factors surface in literature studying ratings of bands at music festivals: teaching experience of the director, quantitative factors such as band size, rehearsal time and budget, and lastly the reliability of contest scores and judging criteria (Groulx 2010.) It is easy to observe and measure factors such as experience, education, and tenure at a particular music program (Beaver, 1973), and a

positive correlation was found between the higher achieving programs and the amount of education a director has according to Dawes (1989), Davis (2000) and Fosse (1965). These authors also observed that as a band director gained more experience, his program's festival ratings improved. In addition, Davis (2000) found a greater interest in musical competition in younger directors rather than more experiences music directors. It was also discovered by Rickels (2008) that the number of days a director rehearsed his band would affect the final ratings received at an adjudicated festival.

Using a test by Hersey and Blanchard, Goodstein (1984 & 1987) gauged the leadership effectiveness of music directors as measured on a self-test. He found many similarities between a group of band directors that had success in their field and a randomly selected group of directors. It was also found that the motivation a music director shows would also correlate to program achievement. Items such as concern for home, parents, ethics, values and security were strong indicators of ensemble performance success (Caimi, 1981). Davis (2000) studied the size of bands and found a positive correlation between the size of the band and the rating it achieved at a music evaluation festival, with larger bands receiving higher ratings. The size of the school a music program belongs to can affect overall scores at music festivals as well (Saul, 1976). However, with respect to sight-reading scores at a concert evaluation festival, Harris (1991) found a very low correlation between score and band size.

Measuring Success in Music Education – Student Influences

According to Washington (2007), the aspects that ultimately affected a music program's overall ratings at music evaluation festivals were those of the school and the students themselves, rather than those of the director. He found that there was a positive correlation between the student's level of musicianship (as measured by the Long-Hoffer Musicianship Test) and the band's overall achievement. Harris (1991) also concluded that the percentage of 11th and 12th grade students in the music program had a positive effect on the band's scores at evaluation festivals. In addition, he found that the amount of 9th grade students in a high school music program negatively affected the band's scores, particularly in sight-reading. The amount of students in a program who took private lessons was the most significant positive factor towards a band's success at evaluation festivals (Washington, 2007). In comparison, students' contributions and decisions towards the music making process of the program had no significance in the way the band performed (Petters, 1976).

Measuring Success in Music Education – Other Influences

Goodstein (1984) and Washington (1987) both found that band budget, as well as the funding sources, were factors contributing to success with respect to concert band festival ratings. Money brought into the program through fundraising gave the strongest positive correlation between budget and ratings, followed by student fees and lastly budget money allotted from the school or district. Over-rehearsing a music program, as described

by Rickels (2008) might lead to a lack of student enthusiasm and therefore a less passionate performance. He also concluded that if directors were using an entire season to prepare only the pieces of music that will be evaluated at festival, perhaps they are performing music that is above the students' ability level. Working on many pieces of music throughout the season will help reinforce basic music fundamentals, develop a better band sound and provide a break from the monotony of practicing the same thing for extended periods (Groulx, 2010). The notion that some directors were good at only certain aspects of music education was dispelled by Dawes (1989) who found no correlation between achievements in marching band ratings versus concert band ratings. He did note, however, that as a director focused more attention on performing only a few pieces of music, a band's sight-reading score would drop at concert evaluation. Rickels (2008) found that as a band attended more festivals, and reviewed the increased commentary and evaluations from the judging panel, it tended to score higher. Burnsed, Sochinski & Hinkle (1983) argued however, that this is more likely due to a reverse causal relationship, where music programs and directors that are already successful will attend more festivals in order to showcase their musical talents. Sheldon (1994) found that students who were preparing music for an adjudicated performance considered the music to be of a better quality than music prepared for a non-adjudicated concert, possibly leading them to work harder in preparation. It was also found that input from adjudicators and exposure to other music programs were seen as reasons to attend music festivals, while drawbacks included limited budget, disorganization at festivals and inconsistencies in the judging community (Sullivan, 2003). A study by Guegold (1989) examined the possibility of judge inconsistence in the

Ohio Music Educators Association (OMEA). Comparing results over a three-year period at the OMEA state finals, he watched for consistency in band's ratings. Although no statistical correlation was found, he did observe that bands attending OMEA state finals had a reasonable chance of receiving a fair and consistent evaluation.

Measuring Success in Music Education – Teaching Styles

A music director's teaching style is described as the way that he or she balances the obligations of teaching and assigns levels of priority to these various aspects of the profession (Groulx, 2010). Such responsibilities might include rehearsing music, teaching basic musical ideas, administrative duties, discipline, making announcements and fundraising. There has been some research into the teaching styles of music educators and the program's quality of performance and festival evaluation ratings. A study by Smith (1999) observed a set of music directors' use of verbal and nonverbal communication. Directors who spoke more about notation, style and rhythm were found to gain higher ratings at evaluation festivals. A study by Bauer (1993) also added that directors who rehearsed concepts such as balance and intonation with their students achieved higher ratings. It was also found that discussions on expression correlated with higher festival ratings than discussion on general notation matters.

A director, who praised his students more directly rather than making general comments towards the group, was found to achieve higher ratings at music evaluation festivals (Groulx, 2010). Also, when a director spoke using verbal imagery rather than

using demonstration or modeling, ratings were negatively affected. Costello (2005) found that directors who self-reported as having good classroom management skills, and felt that their school district provided quality professional development in classroom management had a significant positive correlation with ratings at music festivals. Price (1983) discovered that directors who scored the highest ratings held rehearsals where the students were largely on task, eye contact was constantly made and any non-musical activity was limited to five or six seconds. Also, it was discovered by Yarborough and Madsen (1998) that higher rated music programs rehearsed shorter sections of music during rehearsals, rather than longer passages. In a study by Davis (1998) it was found that as students improved over the course of 40 observed rehearsals, the amount of teacher instruction during those rehearsals was decreased. Gumm (2003) during a study of choir directors' teaching styles found that those who paid closer attention to artistic aspects of the music and used nonverbal communication received higher ratings at evaluation festivals. Teacher-directed classroom styles were found to be more widespread in the music classroom than student-directed styles (Bazan, 2007). While these results were also compared between male and female directors, no impact in differences relating to gender could be found. It was found that younger music educators tend to use the student-directed approach more often. This is believed to be a result of new teachers being accustomed to the more student-centered teaching strategies that are part of a typical teacher education program (Groulx, 2010). This notion is also supported by Hamann (1990) and Spurlock (2002), however they found in general a student-centered classroom led to higher levels of musical success. It was discovered by Teachout (1997) that younger music teachers rank

the importance of student behavior lower than experienced educators. He believes that this may simply stem from a lack of awareness of what is realistically necessary in maintaining an effective and musical classroom environment.

According to Costello (2005) and Davis (1998), teaching style can have an effect on performance outcomes and ratings at music festivals. Their research shows that while student-directed classroom styles are not as common in the music classroom, they are strategies used by more effective, non-music teachers. Gumm (2007) developed a series of eight different teaching styles in an attempt to help teachers understand the outcomes of each style, and when and where to implement them. Brakel (1997) investigated the relationship between director teaching style and program dropout rates, fining no real correlation with respect to any one style. However, he did discover that certain combinations held a positive correlation with dropout rates. Pairings that pointed towards a low degree of student self-direction strategies and higher teacher control tended to increase the dropout rate. In comparison, pairs with greater student freedom, expression and independence showed a lower dropout rate.

Music Education Policy and Law

According to Barresi and Olsen (1992), there is little study that shows the effect educational policy decisions have impacted music education. Years later, Hope (2002) added that the application of policy and its effect on music education is one of the least studied subjects. The No Child Left Behind Act of 2002 integrated the arts as a core subject;

however with most school district's curriculum already in place, this law did little to change what was previously developed or what type of classes were offered (Aguilar, 2011). A study by McIntyre (1990) attempted to bring awareness to legal issues that were part of directing a public school music program. This study found that if a director kept to established educational policies, there was less of a chance that there would be legal controversy. He also believed that a school's administration was an unreliable source when it came to legal issues and that most lawsuits stemmed from decisions that were made without proper thought and consideration. A study by Richmond (1992) researched if the differences between school districts arts curriculum offerings somehow violated the Fourteenth Amendment and the equal protection clause. He concluded that as education is not considered a fundamental right, any differences in offerings would not be a violation of the amendment. However, a case might be made that since a state's constitution mandates education, it could be an amendment violation under state language (Aguilar, 2011).

Pinpointing legal issues that face music educators was the focus of a study by Kerr (2002). The study investigated legislative acts and cases that set legal precedent and laws that affected music educators and the preparation of future music educators. It was concluded that it was in the best interest of music educators to be aware of laws and legal decisions that related to their duties as a professional music educator. A study into the legal right to music education by Heimonen (2006) reflected on the laws of other countries and their different goals, values, sense of justice and traditions. She discovered that while the United States does not directly mention education as a fundamental interest, in Nordic countries students have the right to music at school. In addition, Swedish schools have used

the United Nations principle of a child's best interest to justify music education. Heiminen also believed that internal aims, such as achieving high musical standards, should take priority over external goals such as fame and money. None of these studies, however, have studied the formation of the education laws, or the decision making process in creating those laws (Aguilar, 2011).

Music Policy Formation & Implementation

There is little research into the formation of music education policy and most of the research is in the form of recommendations made by particular organizations (Hope, 1989, 2007). Hope developed policies for music education that suggests thinking of the humanities in the economic context would be the best action to guide how policy is formed in the future. Shuler (2001) suggested ways to draw students towards music in the public schools, how to improve teacher education, and how to build a stronger rational for music education in the schools. Research by Schieb (2006) included policy towards music teacher retention, job satisfaction, managing stress and educator self-identity.

Crone (2002) led a study on the impact that the federal government had on educational policy in the state of New York from 1950-1999. While he was able to classify several different kinds of typically used philosophies, he was unable to show where or how they began or what problems they were making an attempt to solve. Often, certain organizations find themselves in the center of the music education policy debate (Aguilar, 2011). A study by Hoffa (1988) outlined the connection art education has with certain

federal government programs such as the Works Progress Administration, which provided job opportunities for musicians and artists in the 1930's during the Depression. In addition, he outlined several platforms that music educators stand divided on that make arts education more difficult to define and solidify.

The Music Educators National Conference (MENC) was the focus of a study on policy making by Colwell (1994) and Hoffman (1994). It displayed some of the limitations of the organization, such as diverse geological makeup and short terms of office, that make if difficult to create strategic music education plans. Hoffman stated that due to these limitations, MENC would be better served to simply focus on being an advocate in the field of music education. Colwell noted that other organizations, such as the National Education Association, carried more of a greater national voice, yet did not pay particular attention to arts education.

According to Aguilar (2011) there seems to be less research on policy formation than there is on decision-making and the application of music-related policies. McLaughlin (2006) believes that where the policy is being applied, who is executing the policy and how the policy is being implemented is a main focus of policy implementation research. In the case of standards for the arts in the state of Florida, standards were not taken directly from any national association, but rather several, and modified to suit the needs of the state (Lee, 1997). VanPatten (1997) developed a model for implementing national music standards into high school music programs. The program focused on creative musicianship and comprehension of musical skills that supported the national standards. Van Patten provided results that showed national standards could be incorporated successfully into

both performance and non-performance classrooms. Lambourne (2002), in a study of music education in the primary classroom, showed how barriers existed in the implementation of federal and state policies and how other programs that favored state testing received greater attention. Lambourne went on to recommend that further research on music education and its role in brain research was needed in this age of testing and accountability. Music teachers feel that due to their courses being on the edge of school curriculum, their concerns are not being listened to by administrators (Kos, 2007). On the positive side however, he found that if a school's administration found importance in a school's music program, it was less likely to feel the effects of federal and state policies.

The body of published work on implementing national standards in the music classroom consists largely of reflections and informal observations from trade journals in music education according to Fallis (1999) and Snyder (2001). According to Wells (1997) no specific procedure exists on how to align local school curriculum to national standards as there is a concern on making sure that any standard-based curriculum has meaning and depth to teachers in the music classroom. Many of the standards, according to Fallis (1999) were difficult to translate into large performance ensembles, such as teaching composition. Snyder (2001) added that standards such as improvising, composing and music history were the most difficult to cover. Both Snyder and Kerchner (2001) provided suggestions on how to satisfy these standards in a middle school music program by using suggested literature. In a study by Kirkland (1996), an evaluation of music standards used in K-12 music programs was conducted. It was found that while the singing and playing standards were consistently met, composing and improvising was ranked at the lowest proficiency

levels. Gillespie (1998) went on further to suggest that using these standards in the music classroom could only serve to help remove any sense of mystery of the music profession with respect to parents and administrators.

Teacher Evaluations at the National Level

After the federal government enacted the Race to the Top program, arts administrators and teachers have tried to develop a consistent model of student assessment (Perrine, 2013), and recent news articles have demonstrated the burden placed on arts teachers and supervisors. In one example from the state of New York, it was suggested that all music directors rate their students on a one to four scale at both the start, and the end, of the school year. These scores would in turn be used to measure the teacher's effectiveness, and ultimately decide on their related job status, raises and tenure (Winerip, 2012). After deciding that this method did little in the way of evaluating and educating teachers, the state music supervisor said there was simply no way to afford an outside effective and objective set of evaluators or consultants. Others who have been charged with creating and shaping assessment procedures have said that they believed state official did not seem concerned with listening to the input of local teachers or administrators who felt that a good concert performance was a better indicator of the teacher effectiveness than Race to the Top generated paperwork (Cochran-Smith, 2007). This was in direct contrast to a statement from the state's education commissioner, who

argued that music teachers were open to the idea of this new system of accountability (Perrine, 2013).

According to Cochran-Smith (2007) these are the types of issues and teacher accountability approaches that are facing music educators today. She believes that the attempt to associate teacher effectiveness solely with student scores on a standardized test is a hazard in teacher evaluation. Within the context of the federal Race to the Top grant, teacher accountability is seen as an effort to hold teachers responsible for nothing more than the test scores and learning gains of students (Perrine, 2013). This is a method borrowed from the business world, and is referred to as the value-added model. The main idea of this model is tracking student progress over several years using standardized tests and using this information to come to a decision on the effectiveness of an educator (Edgar, 2012). According to Perrine (2013), this has been a process that has developed through continual cycles of educational reform. While the process began at the state level in the 1960's, federal attention towards teacher accountability has its origins much later on in the 1990's movement concerning standards (Abeles, 2010).

In 1994, the reinstitution of Title I of the Elementary and Secondary Education Act demanded testing in the subjects of reading and math, and for states to develop curriculum standards to monitor the progress of students (Goertz & Duffy, 2003). Although it was required that states meet certain proficiency goals on its standardized tests, a schedule was not established as to when this needed to be achieved. Also, according to Shaul and Ganson (2005), Title I defined actions a school district could take if a school did not meet a certain adequate yearly progress (AYP), including supplemental services and school choice.

However, as these provisions were not required, student progress usually had little consequence for either school districts or the teachers within them. In 1994, the Goals 2000: Educate America Act focused on standards in the curriculum rather than assessment of these standards (Perrine, 2013). During this time, national standards for music were developed and functioned as benchmarks for the development of teachers and the rating of students. Eventually the United States Congress passed the Higher Education Act (HEA) in 1998, which required states to make data available about the manner and quality in which teachers were trained (Walsh, 2004). According to Perrine (2013) when the states failed to effectively report under the HEA, the federal government was pressured to address the issue of teacher quality and school reform for itself. It did so with the creation of the No Child Left Behind Act (NCLB) in 2001, which marked a new approach to teacher accountability. Before NCLB, it was assumed that a teacher with proper state certification was competent to teach. However, after NCLB, the federal government stated that state certification was no longer enough, and states and schools were now required to show that yearly progress had been made (Walsh, 2004). Included in this would be statewide student assessment in reading and math across multiple levels beginning in the year 2006. Scores would be separated by race and socioeconomic status, as well as graduation rates (Shaul & Ganson, 2005). By the year 2014, all students were expected to be achieving at a proficient level, or consequences such as loss of finances, student transfer or state takeover of the school might occur (Perrine, 2013).

In addition to the pressure placed on schools and school districts, classroom teachers were also held to increasing demand from the new federal benchmarks, such as

being designated as “highly qualified” by 2006 (Berry, Hoke & Hirsch, 2004). According to Rebel and Hunter (2004), if a teacher was to teach in a core subject, which included the arts, he or she could become “highly qualified” by earning a bachelors degree, gaining state certification and passing a content area test proving competence in his or her subject. However, Berry, Hoke and Hirsch (2004) believe this will not strengthen teacher quality, as almost any teacher, even those with no teaching experience or those who became certified through alternative means, can easily attain the status of “highly qualified”. In addition, there was little connection between the federal government’s definition of “highly qualified” and any measurable teacher practice (Perrine, 2013). No Child Left Behind placed an emphasis on teacher knowledge rather than teaching ability, going so far as to having the Secretary of Education recommend eliminating student teaching requirements, according to Rebel and Hunter (2004) and Goertz and Duffy (2003). They go on to state that even though teacher quality was the center of NCLB, ultimately accountability would fall into the hands of administrators and school principals and was focused on the progress of the school as a whole.

With the introduction of the Race to the Top program in 2009, major changes to teacher accountability occurred (Perrine, 2013). The program was designed to increase the performance of students by having states compete for a block of federal money. Two of the most urgent aspects of the program include the improvement of teacher effectiveness across demographically diverse schools, and the implementation of strict standards in core subjects (Perrine, 2013). In the largest section of the plan, student test scores are linked to teacher performance as well as an evaluation of the teacher’s undergraduate training

program. While the program has shown success in shaping the federal government's relationship with local school districts and how states handle education reform, it remains voluntary, and a few states have declined participation (Hourigan, 2011).

Educator Self Evaluation – Implementing State and National Standards

A study by Wang and Sogin (1997) compared observed time usage of general music teachers with activity times that were self-reported. It was found, in general, that teachers miscalculated the amount of time they were actually spending on each activity. This particular study did not, however, address or analyze any national standards. Byo (1999) discovered that music teachers felt less confident in implementing the standards than was indicated by their training. At the elementary level, music teachers found it most difficult to implement instrumental playing and improvisation standards, while secondary teachers found composing standards the most difficult (Aguilar, 2011). In addition, while teachers found the national standards had merit and value, they were concerned with the amount of instructional time they had to implement them fully.

One such study conducted by Louk (2002) investigated how general music teachers used the standards in their classrooms. The results indicated all of the standards were witnessed during the observed lessons, and there was a strong correlation between those observations and what the teachers self-reported. Orman (2002) compared time used by grade 1-6 music teachers in an attempt to define and categorize the national standards. Orman's research indicated that standards that involved singing and playing were

addressed the most and that while all the standards were eventually addressed the same level of detail and attention was not given to each standard equally. Lastly, Orman noted a lack of sufficient time for music teachers to address each standard to the same degree.

A few studies have addressed how standard implementation might be addressed in specific types of music ensembles. One such study by Scott (1996) attempted to create an assessment sheet for the national standards in sight singing. Scott's results showed that while students with four years of experience were able to meet the benchmarks, students with one year were not able meet the established standards. In another study by Riveire (1997) an attempt was made to find which standards were being used in K-12 string ensembles, and more specifically if the improvisation standard was addressed. Results showed that while teachers had positive attitudes towards improvising in the classroom, they did lack the confidence and skills needed to teach the standard. Teachers felt that the improvisation was typically associated with jazz music, and they were unclear as to how to bring that skill into other performance ensembles. McCurry (1998) conducted a study of elementary-age music students that were members of either a handbell choir, chorus, instrumental ensemble or a general music studies class. Using an evaluation sheet created by the researcher, assessment results showed that students who participated in the handbell choir achieved the highest ratings on six of the nine tested national standards. McCurry suggested that students who participated in a handbell program were able to achieve the benchmarks faster and with greater ability. Skube (2002) conducted a study in an attempt to gather information on how the national standards were being used in secondary instrumental music programs. Results showed that most skills were being fully

implemented into the music programs (performing, evaluating, history, culture and understanding), some were taught to a lesser extent (reading and notation), while composing and improvising were not being taught. A study by Diehl (2007) indicated the level to which music directors were integrating the standards in a concert band setting through a self-report from participants. The results found that listening to and evaluating music was rated the highest, understanding music in relation to history, culture and other disciplines was next and improvisation and composition was rated the lowest. According to the directors who participated in the study, factors that influenced their ability to implement the music standards included such items as school demographics, teacher development, school curriculum and accountability.

An investigation into the use of the national standards was the focus of a study by Schopp (2006). Through a web-based survey, data were collected from high school concert band directors on their implementation of the standards. In addition to the online survey, five schools were visited in person by the researcher. The results showed that there was support for use of the standards overall, but that a lack of time, or a general anxiety about teaching certain standards, kept them from addressing all the benchmarks in the classroom. Younger teachers appeared to have more knowledge of and a greater support of the standards, which Schopp believed was due to teacher education programs recently placing more focus on teaching the national standards. While it is believed that uses of the national standards were becoming more prevalent in the music classroom, Hinckley (1997) suggested that veteran teachers might be less willing to incorporate them, as they are aware that many educational policies and innovations tend to change quickly. Hinckley also

believed that applying any new educational policy in the classroom could take up to 25 years, as new teachers slowly replaced veteran teachers. Abril and Gault (2006) studied to what degree school administrators were aware of the national standards in music in the Massachusetts school system. It was found that while there was an awareness and support of the standards, there was little work done in the way of actually implementing them. Another study by Abril and Gault (2008) looked at the perceptions of school administrators with regard to national music standards. Ranked highest among administrators was music performance while creating and composing ranked the lowest. This study seems to be in line with the perceptions and implementation of the standards by music directors (Aguilar, 2011).

College and university teacher education programs have the task of making the national music standards known to future music educators (Gillespie, 2001). As most college music programs are built on the European model of music conservatories, it might pose a challenge incorporating new standards into the music curriculum (Shuler, 1995). McCaskill (1998) studied the knowledge, attitudes and educational practices of college music education as well as college professors with respect to national standards. Results of the study showed that most professors were aware of the standards and believed they would improve the quality of public school music education. Others went on to say that even though discussion of the standards appeared solely in methods courses, all music professors should be able to address them throughout the curriculum. A survey by Fonder and Eckrich (1999) addressed changes in different areas of music education curriculum in

the universities. While changes were noted in most music course offerings, as predicted by the researchers, most changes took place in the music education sequence.

Teacher Assessment In Florida

Race to the Top (RTTT) has led states to develop a wide variety of methods to measure teacher value (Perrine, 2013), and while states must meet specific outcomes, the direction they chose to arrive there is not defined. Hourigan (2011) asserts that Race to the Top is nothing more than a practical approach to create competition between states, so it is not surprising that states have taken a wide variety of courses to implement RTTT policies, especially when it comes to hard-to-measure subjects such as music and art. While some states have designed teacher evaluation programs without taking into consideration the needs of music teachers, Florida has adapted a model that calls for cooperation between both the policymakers and the classroom teachers (Perrine, 2013). In addition, the Florida Department of Education believes that a main component of Florida's approach is the creation of content standards and a balanced approach to assessment of students. In one case Polk County Schools, through a $20 million dollar grant, created assessment standards for music by creating the RTTT Performing Fine Arts Assessment Project that will serve as a national model (Race to the Top Assessments, 2012).

The Student Success Act of 2011 brought Florida education laws even closer to the goals of the federal Race to the Top program, and since the 2011-2012 school year, 50% of a teacher's yearly evaluation and pay raise has been based on an assessment from their

principal (Perrine, 2013). According to the Review and Approval Checklist for RTTT Teacher Evaluation Systems (2012), the remaining 50% of a teacher's score is based on student results on the Florida Comprehensive Assessment Test (FCAT), and a teacher in a subject such as music will have to depend on the school's total reading and math scores while specific content area tests are being developed. Tenured teachers can now be let go after receiving unsatisfactory assessments for two consecutive years, and multi-year contracts are no longer offered to new teachers who can be dismissed at any time after receiving only one bad review (Race to the Top Assessments, 2012). The elimination of multi-year contracts for all new Florida teachers is a step that goes far beyond the requirements of the Race to the Top program (Perrine, 2013).

According to Pistone, (2012) there are problems with the current models of assessment for hard-to-measure subjects, such as music. While a standardized test might be able to measure a student's knowledge and understanding about the fundamental concepts of music, it in no way can measure whether or not a student can actually perform or compose music. Pistone continues on to say that such a standardized test will also not be able to show if a teacher has success in educating students in performing music as an ensemble.

Music teachers in the state of Florida have suggested that performance events judged by an independent panel of adjudicators are the most appropriate way to test music student achievement, even going so far as to naming such current music festivals Music Performance Assessments (Cochran-Smith, 2007). Critics have argued however that these music assessments offer no baseline pretest and cannot track individual student

achievement or individual student learning. However, a music teacher's rehearsal technique will still most likely be affected by standardized testing as they prepare for concert evaluations (Perrine, 2013). It is not outrageous to also assume that musical performance might suffer as teachers become more focused with test preparation when salary, benefits and job security are at stake. Another suggested approach might be one that is based on a music teacher's portfolio. Instead of using students test scores as 50% of a music teacher's evaluation, a mixture of other performance-based aspects might be used (Winerip, 2012).

CHAPTER 3: METHODOLOGY

Purpose & Background

The purpose of this study was to determine adjudicator reliability and the degree of perceptual influences in the scoring of musical performances by Florida Bandmasters Association adjudicators. This study examined the criteria contained on the Florida Bandmasters Association concert band music performance assessment instrument, and how an adjudicator under a contrasting set of circumstances interprets them, which might affect the outcome of a concert band's final assigned rating. The study examined the possibility that subjective factors such as the reputation of a school music program, reputation of a director, band size, age of a director, or gender of a director that are only observed in a face-to-face evaluation can have an impact of the final rating assessed by the adjudicator at a FBA Music Performance Assessment. Furthermore, it is believed that by identifying any inconsistencies, the Florida Bandmasters Association may be better able to properly prepare judges and enhance the learning experience of the music programs that participate, as well as providing a more standardized and objective evaluation method.

There are a few research publications that focus on some observable elements such as the race of the director, ensemble uniform choice, the directors conducting style or even the stage presence of the musicians and how these components can affect the perception of an ensembles musical performance (Bradley 1972, Elliot 1995/1996). However, there is little to no research that simply tests the reliability of the Music Performance Assessment Ratings Sheets used by the Florida Bandmasters Association during a concert band festival.

There is a need to examine the criteria contained on these assessment instruments, and how an adjudicator under a contrasting set of circumstances interprets them, which might affect the outcome of a concert bands final assigned rating. The results of this study may provide valuable information that could lead to better development of a fair and balanced rating system.

Research Design and Appropriateness

The Kruskal-Wallis Test, developed in 1952, is a nonparametric test. It is used for comparing two or more independent samples where different sample sizes may exist, and the assumptions of an ANOVA are not met (Corder & Foreman, 2009). In an ANOVA, there is an assumption of normally distributed groups and an approximate equal variance for the scores of each of the groups (Dunn, 1964). According to Siegel and Castellan (1988) the Kruskal-Wallis test holds none of these assumptions, however it does assume that population samples drawn are random, each group is independent and the measurement scale for each group is at least ordinal. In rejecting the null hypothesis of this test, one sample will statistically overshadow at least one of the other samples. The test did not identify where this dominance occurred and specific sample pairs were analyzed in post-hoc testing to find where the differences occurred (Spurrier, 2003). Any statistical significance found was followed by a Mann-Whitney test between groups to determine where the differences existed.

The chi-square test for association, also referred to as the chi-square test of

independence, was used to test to what degree the four groups of adjudicators scores were statistically associated or independent (Lund, 2013). Effect size for the post-hoc comparisons was calculated using standard effect size guidelines (Yatani, 2014). The Friedman test was used to determine if the adjudicator's medians for the three sub-captions, as well at the criteria contained in each of the three sub-captions, differed within the population (Lowry, 2015). Pairwise comparisons were made between the sub-captions using a Wilcoxon test, but were not done for the sub-caption criteria, as that was not within the scope of this study.

Setting

After receiving approval of the International Review Board (see APPENDIX A) research took place through a webpage where the participants could listen to the musical excerpts to be evaluated and fill out an online evaluation form. The music excerpts and online forms could be accessed from any public or private computer with an Internet connection. This could be done at the participant's leisure in any setting.

Consent Process

Consent was obtained from all participants in this study. Consent was obtained by providing the participants with a copy of University of Central Florida form "HRP-502a: Consent – Adult". The principle investigator followed form "SOP: Informed Consent Process

for Research (HRP-090)". Consent of the participants was document by their choice to participate in the study and by answering the provided questioners (see APPENDIX B).

As the research involved minimal risk to the participants, written documentation of consent was not required and signatures were not obtained. Participants received a copy of the consent form for their records. As per University of Central Florida form "HRP-411 CHECKLIST: Criteria for Waiver of Written Documentation of Consent", written documentation was not required as the written script of the information included all required and appropriate additional elements of consent. These elements included:

- That the study involved research.
- The purposes of the research study.
- The expected length of the subject's involvement in the study.
- Participation in the study was voluntary.
- The procedures of the study.
- Any risks or discomforts to the participant.
- Contact information of the research team for questions, concerns or complaints about the research.
- Contact information outside of the research team for questions, concerns, complaints, questions about subjects' rights, information, or to offer additional input.
- Contact information in the event of a research-related injury to the subject.
- Refusing to participate will not invoke any penalty or loss of benefits and the subject may terminate participation at any time.

- The research involved no more than minimal risk to participants.
- The research procedures did not require written consent.
- Written information describing the research was provided to the participant or their legally authorized representative.

Participant Process

When participating in this study, the subject was asked to listen to a musical performance, and evaluate the performance using an online questioner. From any personal computer, the participants were directed to a website that guided them through the process. On this website they found five musical examples, and five corresponding links to answer questions about those musical examples (see Appendix A). The study participants did not have to complete all the surveys in one sitting, and had 30 days to complete all of them. Each survey was to be completed only once per participant. Participants were instructed to:

1. Click on a musical excerpt to listen to it directly, or download it, from the provided webpage.

2. Click on the corresponding link that took them to an online survey where they answered questions about the performance they just heard. As the survey will open in a separate browser window, participants were able to listen to the musical performance and respond to the survey simultaneously, as might be done during a typical concert band evaluation. Participants were reminded to press "submit" at the end of each survey before

closing the window.

3. Repeat the process for the remaining musical excerpts and corresponding surveys (a total of five).

4. Click on the last survey link to answer some general questions about their professional musical beliefs and asked them to rate the criteria used in the Florida Bandmasters Association's concert band MPA (see Appendix B).

Withdrawal of Participants

If a randomly selected participant was found to have any prior attachment to or affiliation with the music performances being evaluated, they may be withdrawn from the study. The participant would be notified in writing that they have been removed from the study. Any data collected from a participant that was withdrawn (either voluntarily, or without their consent) were not included in the study, and a new participant was selected in their place.

Risks, Benefits and Participant Privacy

Participants were not required to travel anywhere public to partake in the study, so there were no applicable privacy interests or concerns. All portions of the study could be done in a private residence if desired. There was no direct benefit to the participants and potential risks may have included:

- Loss of time, as time to complete the evaluation will take approximately two hours.
- Mental Fatigue.
- Frustration.

Participants and Selection Process

Samples of five audio recordings were collected. One was chosen from each of the top five Florida public school concert band directors who have the highest frequency of superior ratings at the Florida Bandmasters Association Music Performance assessment. Any extra silence or metadata was removed from the recordings before posting them for use in the study.

There are currently 236 concert-band certified Florida Bandmaster Association music judges. These adjudicators are certified to judge in any one of the 21 FBA districts across the state. A sample of ten adjudicators was selected from this population. Ten non-certified FBA adjudicators from the state of Florida, as well as ten certified adjudicators from outside of Florida were also selected. A website link was sent to each of the 30 adjudicators which included for their review:

- Five separate MP3 recordings of state-level, superior rated concert band performances.
- A corresponding link to an online survey for each digital recording, which contained the evaluation criteria to be used.

- An online survey that asked participants to rank the sub-captions, and the criteria within each sub-caption, that are found on the Florida Bandmasters Association concert band MPA assessment instrument (see APPENDIX E).

Information such as final ratings, and other musical elements were collected and compared to the information and ratings given by the judges at the initial performance. In order to lessen any potential risks to participants each participant was given 30 days to complete the evaluation. The research period will be approximately 45 days from participant selection until primary analysis.

Data on a participant's membership, associations and qualifications may be collected when selecting participants. Source records used to collect data about the participants included:

- The Florida Bandmasters Association list of certified Concert Band Judges.
- The Florida Bandmasters Association member list.
- The Central States Judging Association member list.

Provisions to Maintain the Confidentiality of Data

Identifiable data were not linked to participants who contributed to the study by answering the survey questions. Surveys did not ask for any additional identifiable data from the participants. Identifiable records of the participants was not collected or used in the reporting of data. Data will be stored electronically locally and backed up using an off-site cloud server, both under password protection. Only the principle investigator will have

access to the password-protected data. Data will be stored for five (5) years as per University of Central Florida policy.

Research Questions

1. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind certified FBA adjudicators?

2. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-certified FBA adjudicators?

3. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-local certified adjudicators?

4. How do adjudicators rank the importance of the three major sub-captions and the criteria within each sub-caption?

Hypothesis

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: At least one mean score is not statistically equal.

Data Analysis

All data were transmitted and transported electronically through the use of a website, online media player and electronic questionnaires. The principle investigator was responsible for collection and management of the data. For management and control purposes, data were automatically populated into a spreadsheet directly from the online form. These data were then electronically transferred to SPSS version 21.0 software for analysis.

Table 1 below provides an outline of the research questions, independent and dependent variables, data sources and the methods of data analysis.

Table 1: Research Questions, Variables, Data and Analysis

RESEARCH QUESTIONS	VARIABLES	DATA SOURCES	METHODS OF ANALYSIS
What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind certified FBA adjudicators?	Independent: *Adjudicator Group* Dependent: *Total Score*	FBA Concert Band MPA Performances Online Participant Surveys	Mann-Whitney Test Chi-Square Test of Independence Effect Size
What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-certified FBA adjudicators?	Independent: *Adjudicator Group* Dependent: *Total Score*	FBA Concert Band MPA Performances Online Participant Surveys	Mann-Whitney Test Chi-Square Test of Independence Effect Size
What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-local certified adjudicators?	Independent: *Adjudicator Group* Dependent: *Total Score*	FBA Concert Band MPA Performances Online Participant Surveys	Mann-Whitney Test Chi-Square Test of Independence Effect Size
How do adjudicators rank the importance of the three major sub-captions and the criteria within each sub-caption?	Independent: *Adjudicators* Dependent: *Sub-Caption Rakings*	Online Participant Survey	Friedman Test Wilcoxon Test

Validation of the Survey Instruments

Validity of any survey instrument can be separated into four parts, including face validity, content validity, construct validity and criterion-related validity (Cozby, 2009). According to Holden (2010) a survey instrument can have face validity if, simply stated, it appears that it will measure what it intends to measure. Content validity refers to the amount to which a survey represents a particular area of study and agreed upon by experts in a given field. Construct validity refers to the degree to which a survey measures what it proposes to measure. Here, a statistical analysis of the tests internal structure is required (Lawshe, 1975). Lastly, criterion-related validity shows a correlation between a survey instrument and other similar tests that are already considered valid (Cozby, 2009).

A pilot study was conducted to test the validity of the "Musical Example Evaluation Form" and "Order of Importance" survey. Twenty participants were selected to take the survey and to leave any feedback for the researcher, identifying any difficulties they may have encountered. Participants included educators, adjudicators and field experts that were a subset of the study participants selected through random assignment to support internal validity. The participants, with respect to the "Order of Importance" survey, agreed upon face validity and content validity of the instrument. Cronbach's Alpha was used to test the reliability of the "Musical Example Evaluation Form" survey instrument. A reliability coefficient of .830 was obtained (Table 2) indicating a high internal consistency in the responses.

Table 2: Pilot Study, Cronbach's Alpha Test Statistic

Cronbach's Alpha	N of Items
.830	4

Even though Table 3 shows a slight increase in Cronbach's Alpha with the deletion of one item (TechPrep), that items was not deleted as the instrument tested is currently in use by the Florida Bandmasters Association, and previous data have already been collected on that instrument with that item included.

Table 3: Pilot Study, Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
PerFund	4.40	2.147	.605	.809
TechPrep	4.55	2.155	.455	.879
MusicEff	4.45	1.734	.757	.737
Final	4.45	1.839	.870	.695

CHAPTER 4: RESULTS

Introduction

The Florida Bandmasters Association is a professional organization and the governing body of all middle and high school music programs in the state of Florida. Its concern is the development and promotion of public school music programs by providing in-service opportunities through conferences, clinics and what are called Music Performance Assessments (MPA). Music Performance Assessments are performances held several times a year for middle school and high school Marching Bands, Jazz Bands, and even individual solo performers. However, probably the most important and highly regarded of all these evaluations is the annual Concert Band Music Performance Assessment.

At a Florida Bandmasters Association Concert Band MPA, a set of three certified FBA concert band adjudicators evaluate a band's live performance and each assigns the band a rating using a Likert-type scale of 1 (superior), 2 (excellent) 3 (good), 4 (fair) and 5 (poor). The adjudicator also records audio commentary while the band performs to help the participants understand how their performance compares to a set of musical standards. The judge notates any additional comments and the band's rating on an official evaluation instrument, referred to as the "sheet". The sheet contains the criteria the judge must use to arrive at the final rating. Ultimately it is the adjudicator's interpretation of the musical performance that determines what final rating is given. Bands who receive a superior rating from all three judges at the local, district-level performance can elect to have their

band evaluated again, a few months later, at the state-level concert band evaluation. If that band once again receives superior ratings from the entire panel of three judges at the state level, the recording of that performance is placed in a resource library maintained by Florida Bandmasters Association. The recordings in this library are meant to serve as a guide and reference for directors to model when performing the same piece or style of music. An FBA member can request a copy of any recording in this library to use as reference.

Purpose of the Study

The purpose of this study was to determine adjudicator reliability and the degree of perceptual influences in the scoring of musical performances by Florida Bandmasters Association adjudicators. Independent adjudicators from three different populations were asked to evaluate a set of performance recordings obtained from the FBA resource library, and provide a rating for each of those recordings using the FBA concert band assessment sheet. Those ratings were compared to the ratings given by FBA certified concert band adjudicators at the face-to-face performance, to determine whether there was a statistically significant difference in scores given between adjudicator groups. Lastly, this study examined the sub-captions and sub-caption criteria contained on the Florida Bandmasters Association concert band music sheet, and how adjudicators rank their importance when evaluating a musical presentation.

The possibly of perceptual distortions in scores assessed by adjudicators at the

annual concert band Music Performance Assessment was explored. In addition, it was considered whether factors that are only observed in a face-to-face evaluation such as the reputation of a school music program, reputation of a director, band size, age of a director, or gender of a director, might have an impact on the final scores assessed. A few research publications have focused on some observable elements such as the race of the director, ensemble uniform choice, the director's conducting style or even the stage presence of the musicians and how these components can affect the perception of an ensemble's musical performance (Bradley 1972, Elliot 1995/1996). However, there is little to no research that simply tests the reliability of the Music Performance Assessment ratings sheets used by the Florida Bandmasters Association during a concert band festival.

In this chapter, a description and justification of the selected statistical tests will be discussed. Next, the findings will be presented including normality, statistical significance, association and effect size. Finally, a summary of the results will conclude the information presented. It is believed that by identifying any inconsistencies, the Florida Bandmasters Association may be better able to properly prepare judges and enhance the learning experience of the music programs that participate, as well as providing a more standardized and objective evaluation method.

Population and Samples

The sample of audio recordings selected from the Florida Bandmasters Association resource library were from the top-five Florida public-school concert band directors who

have the highest frequency of superior ratings the FBA concert band Music Performance Assessment.

There are currently 236 concert-band certified Florida Bandmaster Association music judges. These adjudicators are certified to judge in any one of the 21 FBA districts across the state. A sample of adjudicators (*n*=10) was selected from this population. A sample of non-certified FBA adjudicators (*n*=10) from the state of Florida, as well as certified adjudicators from outside of Florida (*n*=10), was also selected. Each adjudicator was asked to evaluate the selected recordings using the Florida Bandmasters Association Concert Band MPA assessment sheet. Those scores were compared to the scores assessed by the panel of face-to-face certified FBA concert band adjudicators (*n*=6) that evaluated the original live performance.

The assessment sheet contained three sub-captions (Technical Preparation, Musical Effect and Performance Fundamentals) that the judge was asked to consider and rate using a Likert-type scale with a score of 1 (superior) being the highest score and 5 (poor) being the lowest. The three sub-captions were then tallied to arrive at an average final rating of 1 to 5 for each of the performances. All five scores, one for each recorded performance, were then added together to arrive at a total score between 5 and 25 for each adjudicator. The total scores from each independent group were collected and compared to the total scores given by the panel of Certified FBA concert band adjudicators at the initial face-to-face evaluations.

A final survey asked the adjudicators to rank the sub-captions, and the criteria within each sub-caption, that are found on the Florida Bandmasters Association Concert

Band Music Performance Assessment instrument and the rankings were examined through descriptive statistics.

Test Selection

The Kruskal-Wallis Test, developed in 1952, is a nonparametric test. It is used for comparing two or more independent samples where different sample sizes may exist, and the assumptions of an ANOVA are not met (Corder & Foreman, 2009). According to Siegel and Castellan (1988), in order to utilize a Kruskal-Wallis test, four assumptions must exist; a single dependent variable that is measured at an ordinal or continuous level, an independent variable consisting of at least two categorical groups, independence of observations and data that are not normally distributed.

In rejecting the null hypothesis of this test, one sample statistically overshadowed at least one of the other samples. The test did not identify where this dominance occurred and specific sample pairs were analyzed in post-hoc testing to find where the differences occurred (Spurrier, 2003). Any statistical significance found was followed by a Mann-Whitney test between groups to determine where the differences existed. The assumptions of a Mann-Whitney test include a dependent variable measured at the ordinal or continuous level, an independent variable with two categorical groups, independence of observations and non-normally distributed data (Siegel & Castellan, 1988).

The chi-square test for association, also referred to as the chi-square test of

independence, is used to test to what degree two variables are statistically associated or independent. Although ordinal data can be tested, this assessment will lose any information that is gathered by knowing the order or rankings of the scores. In addition, even as this study contained both dependent and independent variables, the chi-square test for association did not distinguish between them (Lund, 2013). Effect size for the post-hoc comparisons was calculated using $r = \frac{Z}{\sqrt{2N}}$ where N was the total number of samples contained in the test (Yatani, 2014). Standard guideline values for small (0.1), medium (0.3) and large (0.5) effect sizes according to Cohen (1988) were used.

Lastly, the Friedman test was used to determine if the adjudicator medians for the three sub-captions, as well at the criteria contained in each of the three sub-captions, differed within the population. The Friedman test is a non-parametric alternative for use when the data are not normally distributed and where repeated measures for each subject use ranked ordering (Lowry, 2015). Pairwise comparisons were then made between the sub-captions using a Wilcoxon test, but were not done for the sub-caption criteria, as that was not within the scope of this study.

Findings

Before the research questions could be properly addressed, the assumption of normality for the dependent variable (total score) was measured by a Shapiro-Wilk's test for each of the blind adjudicator groups. The test reviled that normality was met for Certified FBA Adjudicators and Certified Non-Local Adjudicators (p>.01). However

normality was not met for Non-Certified FBA Adjudicators (p<.01) as shown in Table 4.

Table 4: Test of Normality For Each Independent Group of Blind Adjudicators[a]

	Group	Kolmogorov-Smirnov[b]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Total	Certified FBA	.200	10	.200*	.893	10	.185
	Non Certified FBA	.312	10	.006	.749	10	.003
	Certified Non Local	.191	10	.200*	.947	10	.627

*. This is a lower bound of the true significance.

a. Total is constant when Group = Face to Face. It has been omitted.

b. Lilliefors Significance Correction

As the data were not normally distributed, a Kruskal-Wallis Analysis of Variance was used to test the null hypothesis that there was no difference in total score between the four independent groups of adjudicators at a significance level of p=.01. The Kruskal-Wallis Test (Table 5) found a statistically significant difference in total score between the groups of adjudicators (χ^2 =20.97, df=3, p<.01), but it did not indicate between which specific groups of adjudicators the differences occurred.

Mean ranks were calculated for each group. The mean ranks in order from greatest to least were Certified Non-Local (26.4), Certified FBA (23.3), Non-Certified FBA (14.5) and Face-to-Face Adjudicators (4.0). This showed that there was a noteworthy difference between the largest and smallest mean rank and supported the conclusion that there were statistically significant differences in the scores assessed between groups (Table 6).

Table 5: Kruskal-Wallis Test Statistics For Scores Between Adjudicator Groups[a,b]

	Total
Chi-Square	20.969
df	3
Asymp. Sig.	.000

a. Kruskal Wallis Test
b. Grouping Variable: Group

Table 6: Mean Rank Scores of Each Adjudicator Group

	Group	N	Mean Rank
Total	Face to Face	6	4.00
	Certified FBA	10	23.30
	Non Certified FBA	10	14.50
	Certified Non Local	10	26.40
	Total	36	

As it has been demonstrated above that there is a statistically significant difference in scores between the independent groups of adjudicators, the appropriate post-hoc testing was selected and utilized in an attempt to uncover where, and to what degree, those differences in scoring existed, thus addressing the research questions that follow.

Research Question 1

1. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind certified FBA adjudicators?

To answer this question, a Mann-Whitney post-hoc test was performed to determine if there was a statistical difference in total score between face-to-face and blind certified FBA adjudicators at a significance level of p=.01. Each of the blind certified FBA adjudicators (n=10) evaluated the five audio recordings using the Likert-type scale of 1 to 5 and a total score for each adjudicator (between 5 and 25) was calculated. Those scores were compared to the scores already given by face-to-face adjudicators (n=6) during the live performance. The test results in Tables 7 and 8 show that total scores given by face-to-face certified FBA adjudicators (mean rank = 3.5) and blind certified FBA concert band adjudicators (mean rank = 11.5) were in fact statistically significantly different (z=-3.357, p<.01).

Table 7: Mann-Whitney Test Statistics, Face-to-Face and Certified FBA Adjudicators[a]

	Total
Mann-Whitney U	.000
Wilcoxon W	21.000
Z	-3.357
Asymp. Sig. (2-tailed)	.001
Exact Sig. [2*(1-tailed Sig.)]	.000[b]

a. Grouping Variable: Group
b. Not corrected for ties.

Table 8: Face-to-Face and Certified FBA Adjudicator Mean Ranks

	Group	N	Mean Rank	Sum of Ranks
Total	Face to Face	6	3.50	21.00
	Certified FBA	10	11.50	115.00
	Total	16		

A chi-square test for association was used to examine to what degree an adjudicator's total score was statistically associated or independent from the adjudicator's group. A pairwise comparison found a perfect association between group and total score with Cramer's *V* reported as 1.0 (Table 9), indicating that the total score assessed by an adjudicator is completely dependent on which group they represent. Further, effect size value calculated at *r*=.59 strengthened the conclusion that there is a strong and significance difference in the total scores assessed between face-to-face FBA adjudicators and blind certified FBA adjudicators.

Table 9: Face-to-Face and Certified FBA Adjudicator Pairwise Comparison

		Value	Approx. Sig.
Nominal by Nominal	Phi	1.000	.014
	Cramer's V	1.000	.014
N of Valid Cases		16	

In addition, Figure 1 visually illustrates the differences in total scores assessed by face-to-face certified FBA adjudicators and blind certified FBA adjudicators. It clearly shows that in not one single case did a blind adjudicator give the same score as a face-to-face adjudicator.

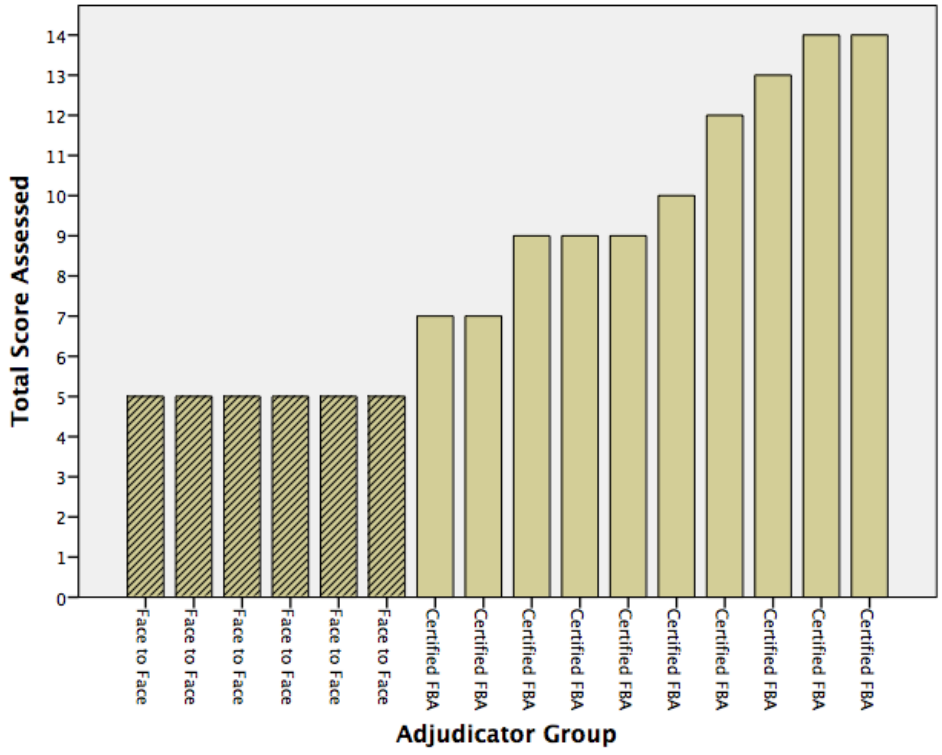


Figure 1: Scores Assessed by Face-to-Face and Certified FBA Adjudicators

Research Question 2

2. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-certified FBA adjudicators?

Again, a Mann-Whitney post-hoc test was performed to determine if there was a statistical difference in total score between face-to-face and blind non-certified FBA adjudicators at a significance level of p=.01. Each of the blind non-certified FBA adjudicators (n=10) evaluated the five audio recordings using the Likert-type scale of 1 to 5 and a total score for each adjudicator (between 5 and 25) was calculated. Those scores were compared to the scores already given by face-to-face adjudicators (n=6) during the live performance. The test results in Tables 10 and 11 show that total scores given by face-to-face certified FBA adjudicators (mean rank = 4.0) and blind non-certified FBA concert band adjudicators (mean rank = 11.2) were once again statistically significantly different (z=-3.107, p<.01).

Table 10: Mann-Whitney Test, Face-to-Face and Non-Certified FBA Adjudicators[a]

	Total
Mann-Whitney U	3.000
Wilcoxon W	24.000
Z	-3.107
Asymp. Sig. (2-tailed)	.002
Exact Sig. [2*(1-tailed Sig.)]	.002[b]

a. Grouping Variable: Group
b. Not corrected for ties.

Table 11: Face-to-Face and Non-Certified FBA Adjudicator Mean Ranks

	Group	N	Mean Rank	Sum of Ranks
Total	Face to Face	6	4.00	24.00
	Non Certified FBA	10	11.20	112.00
	Total	16		

A chi-square test for association was used to examine to what degree an adjudicator's total score was statistically associated or independent from the adjudicator's group. A pairwise comparison found a strong association between group and total score with Cramer's *V* reported as .88 (Table 12), indicating that the total score assessed by an adjudicator is strongly, but not completely dependent on which group they represent. Further, effect size value calculated at r=.55 again strengthened the conclusion that there is a strong and significance difference in the total scores assessed between face-to-face FBA adjudicators and blind non-certified FBA adjudicators.

Table 12: Face-to-Face and Non-Certified FBA Pairwise Comparison

		Value	Approx. Sig.
Nominal by Nominal	Phi	.878	.030
	Cramer's V	.878	.030
N of Valid Cases		16	

In addition, Figure 2 visually illustrates the differences in total scores assessed by face-to-face certified FBA adjudicators and blind non-certified FBA adjudicators. In this case, only one blind adjudicator gave the same total score as a face-to-face adjudicator. Five blind adjudicators all gave the same total score of seven (7), showing a bit more agreement between members of this group.

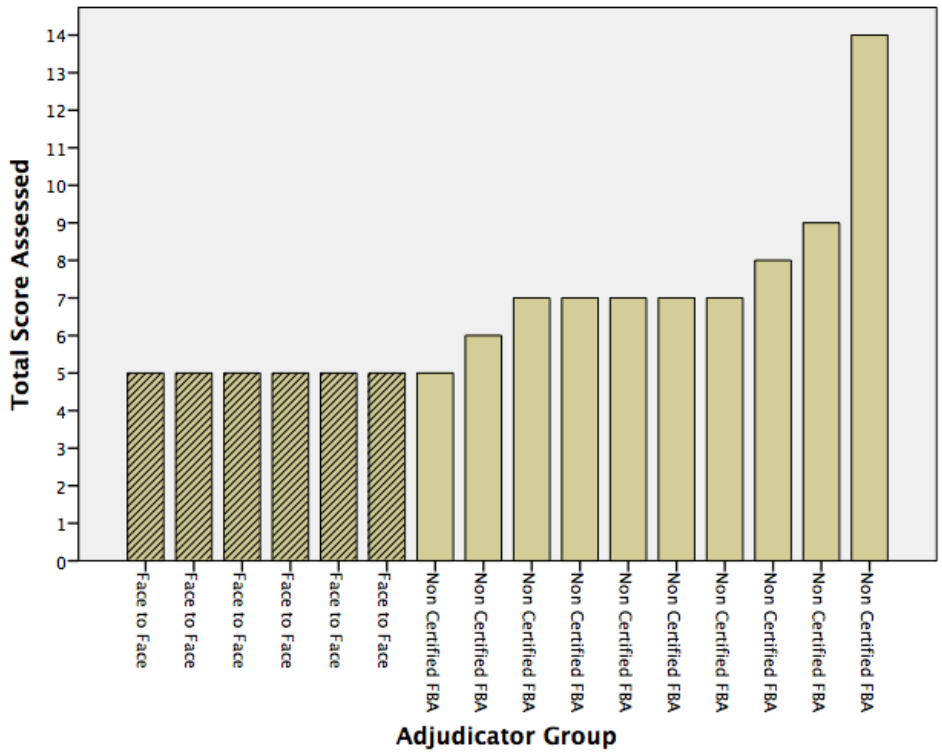


Figure 2: Scores Assessed by Face-to-Face and Non-Certified FBA Adjudicators

Research Question 3

3. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-local certified adjudicators?

Lastly, a final Mann-Whitney post-hoc test was performed to determine if there was a statistical difference in total score between face-to-face and certified non-local adjudicators at a significance level of p=.01. Each of the blind non-local adjudicators (n=10) evaluated the five audio recordings using the Likert-type scale of 1 to 5 and a total score for

each adjudicator (between 5 and 25) was calculated. Those scores were compared to the scores already assessed by face-to-face adjudicators (n=6) during the live performance. Again, the test results in Table 13 and 14 show that total scores given by face-to-face certified FBA adjudicators (mean rank = 3.5) and blind non-local adjudicators (mean rank = 11.5) were statistically significantly different (z=-3.357, p<.01).

Table 13: Mann-Whitney Test, Face-to-Face and Certified Non-Local Adjudicators[a]

	Total
Mann-Whitney U	.000
Wilcoxon W	21.000
Z	-3.357
Asymp. Sig. (2-tailed)	.001
Exact Sig. [2*(1-tailed Sig.)]	.000[b]

a. Grouping Variable: Group
b. Not corrected for ties.

Table 14: Face-to-Face and Certified Non-Local Mean Ranks

Ranks

	Group	N	Mean Rank	Sum of Ranks
Total	Face to Face	6	3.50	21.00
	Certified Non Local	10	11.50	115.00
	Total	16		

A chi-square test for association was used to examine to what degree an adjudicator's total score was statistically associated or independent from the adjudicator's group. A pairwise comparison found another perfect association between group and total

score with Cramer's *V* reported as 1.0 (Table 15), indicating that once again the total score assessed by an adjudicator is completely dependent on which group they represent. Effect size value was calculated at *r*=.59 and further strengthened the conclusion that there is a strong and significance difference in the total scores assessed between face-to-face FBA adjudicators and blind non-local adjudicators.

Table 15: Face-to-Face and Certified Non-Local Pairwise Comparison

		Value	Approx. Sig.
Nominal by Nominal	Phi	1.000	.014
	Cramer's V	1.000	.014
N of Valid Cases		16	

Figure 3 again visually illustrates the differences in total scores assessed by face-to-face certified FBA adjudicators and blind non-local adjudicators. It clearly shows that in not one single case did a blind non-local adjudicator give the same score as a face-to-face certified FBA adjudicator.

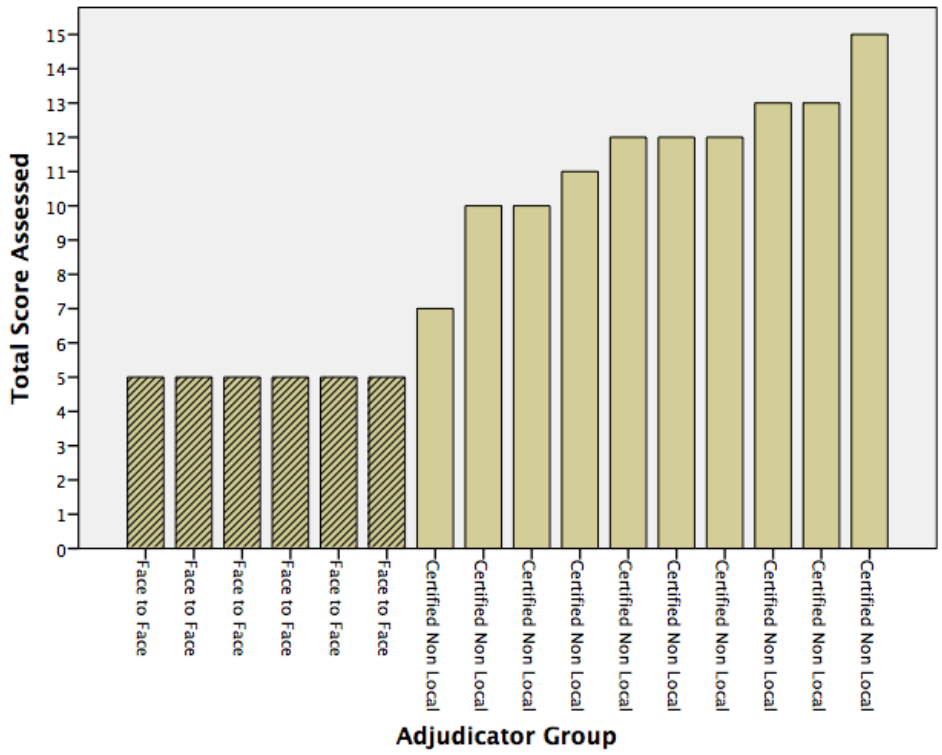


Figure 3: Scores Assessed by Face-to-Face and Certified Non-Local Adjudicators

Research Question 4

4. How do adjudicators rank the importance of the three major sub-captions and the criteria within each sub-caption?

When adjudicating a performance at a Florida Bandmasters Association Music Performance Assessment, judges are provided with ratings sheets for the event to be evaluated. These sheets contain the standards an adjudicator is to use when evaluating a performance and assigning a score. The rating sheet for a Concert Band MPA is divided into three captions: Performance Fundamentals, Technical Preparation and Musical Effect.

Using the same Likert-type scale of 1 (superior), 2 (excellent), 3 (good), 4 (fair) and 5 (poor) the adjudicator rates the performance using each sub-caption and then scores are tallied to arrive at a final rating.

Regardless of group, each blind adjudicator participating in the study (*n*=30) was asked to rank, in order of importance, the three sub-captions contained on the FBA Concert Band MPA assessment sheet. A Freidman test was conducted to test for differences in medians among adjudicators, indicating how the adjudicators ranked the importance of each sub-caption. Performance Fundamentals (median=1.0) was considered the most important by the adjudicators, Technical Preparation (median=2.0) was ranked 2nd and Musical Effect (median=2.5) was found to be considered least important of the three sub-captions as shown in Table 16.

Table 16: Sub-Caption Medians

		PerfFund	TechPrep	MusEff
N	Valid	30	30	30
	Missing	0	0	0
Mean		1.43	2.40	2.17
Median		1.00	2.00	2.50
Mode		1	2[a]	3

a. Multiple modes exist. The smallest value is shown

In addition, the results of the Friedman test were statistically significant (χ^2 =15.27, *df*=2, *p*<.01) indicating that there are in fact significant differences in adjudicators' ranking of sub-captions on the FBA concert band Music Performance Assessment sheet. Kendall's

coefficient of concordance reported at .25 suggested a medium difference in ranking among the three sub-captions as shown in Table 17.

Table 17: Friedman and Kendall's W Test Statistics between Sub-Captions

N	30
Kendall's W[a]	.254
Chi-Square	15.267
df	2
Asymp. Sig.	.000

a. Kendall's Coefficient of Concordance

A follow up pairwise comparison was made through a Wilcoxon test at the p=.01 significance level. The concern for Performance Fundamental (μ=1.43, sd=.568) was statistically greater than that of Technical Preparation (μ=2.40, sd=.621, p<.01) as well as Musical Effect (μ=2.17, sd=.913, p<.01). However, the concern for Technical Preparation did not differ significantly from Musical Effect (p>.01) as shown in Table 18.

Table 18: Wilcoxon[a] Test Statistics Between Sub-Captions

	TechPrep - PerfFund	MusEff - PerfFund	MusEff - TechPrep
Z	-4.288[b]	-2.811[b]	-1.121[c]
Asymp. Sig. (2-tailed)	.000	.005	.262

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

c. Based on positive ranks.

Each sub-caption that is contained on the adjudicated sheet contains a set of specific criteria that the judges are encouraged to consider. While the judge does not necessary have to rate or rank any of the criteria specifically during the performance, then can indicate if a band or performer was noticeably good or inconsistent in any of those areas. Each blind adjudicator in this study (*n*=30) was also asked to rank, in order of importance, the criteria contained within each sub-caption found on the assessment sheet.

The set of criteria contained in the Performance Fundamentals sub-caption include: Tone, Intonation, Balance, Blend, Sonority and Articulation. Tone was given the highest ranking by adjudicators (μ=1.63, *sd*=1.07) and ranked highest overall in 19 cases (*n*=30). Ranked lowest of the criteria was Articulation (μ=5.27, *sd*=1.11) placing lowest of the six criteria in 19 cases, 63.3% of the time (see Table 19).

Table 19: Performance Fundamentals Criteria

		Tone	Int.	Bal.	Blend	Son.	Art.
N	Valid	30	30	30	30	30	30
	Missing	0	0	0	0	0	0
Mean		1.63	2.53	4.10	4.03	3.43	5.27
Median		1.00	2.50	4.00	4.00	3.00	6.00
Mode		1	3	4	4	2	6

The set of criteria contained in the Technical Preparation sub-caption include: Note Accuracy, Rhythmic Accuracy, Precision, Entrances, Releases, Interpretation, Clarity, Technique, Pulse, Dynamics and Transitions. Note Accuracy was given the highest ranking by adjudicators (μ=1.73, *sd*=1.34) and ranked highest overall in 18 cases (*n*=30). Ranked

lowest of the criteria was Transitions (μ=9.73, *sd*=1.72) placing lowest of the eleven criteria in 13 cases, 43.3% of the time (see Table 20).

Table 20: Technical Preparation Criteria

		Note	Rhy.	Prec.	Ent.	Rel.	Intrp.	Clar.	Tech.	Pulse	Dyn.	Trans.
N	Valid	30	30	30	30	30	30	30	30	30	30	30
	Missing	0	0	0	0	0	0	0	0	0	0	0
Mean		1.73	3.07	5.70	5.43	7.03	7.13	6.57	5.80	6.60	6.73	9.73
Median		1.00	2.00	5.00	6.00	7.00	8.00	6.50	6.00	6.50	6.50	10.00
Mode		1	2	5	7	6	7[a]	5	1[a]	3[a]	4	11

a. Multiple modes exist. The smallest value is shown

The set of criteria contained in the Musical Effect sub-caption include: Expression, Shaping, Style, Interpretation, Phrasing, Tempo and Dynamics. Style was given the highest ranking by adjudicators (μ=3.03, *sd*=1.77) and ranked highest overall in 7 cases (*n*=30). Ranked lowest of the criteria was Tempo (μ=5.60, *sd*=1.94) placing lowest of the seven criteria in 14 cases, 46.7% of the time (see Table 21).

Table 21: Musical Effect Criteria

		Exp	Shap	Style	Intrp	Phras	Tempo	Dyn
N	Valid	30	30	30	30	30	30	30
	Missing	0	0	0	0	0	0	0
Mean		3.57	3.80	3.03	4.33	3.30	5.60	4.37
Median		3.00	3.00	3.00	4.00	3.00	6.00	5.00
Mode		1	3	1[a]	4	3	7	5

a. Multiple modes exist. The smallest value is shown

Summary

The purpose of this study was to determine adjudicator reliability and the degree of perceptual influences in the scoring of musical performances by Florida Bandmasters Association adjudicators. Data gathered from the evaluations of the musical samples, as well as from the adjudicator's music criteria order of importance survey, were collected and presented. A statistical analysis of each of the four research questions was performed and the results outlined and reported using narrative, tables and figures where applicable. The results of the tests showed that in almost every individual case, blind adjudicators rated the recorded musical performances lower in quality than certified FBA concert band adjudicators did at face-to-face performances. This held true regardless of which group the blind adjudicators were associated with; either certified FBA adjudicators, non-certified FBA adjudicators or certified non-local adjudicators. This assumption was strengthened by the fact that even judges from the same population of certified concert band FBA adjudicators were in disagreement on total score, as the blind group rated the musical performances lower in quality than any of the face-to-face adjudicators did. In only one instance was the total score assessed between any face-to-face and a blind adjudicator equal. In addition, non-local adjudicator scores skewed the highest (and therefore lowest in quality) of any blind group.

Another noteworthy piece of information was that scores given by both of the blind certified adjudicator groups were statistically equal. This might suggest that proper training through membership in a professional judges association leads to more accurate

and consistent scoring between adjudicators across multiple performances.

The results of the adjudicators' ranking of the sub-captions contained on the FBA concert band assessment sheet showed that Performance Fundamentals were rated higher than both Technical Preparation and Musical Effect. However, statistically Technical Preparation and Musical Effect were found to be no different. Of the sub-caption criteria, Tone, Note Accuracy and Style were ranked the most important by judges, while Articulation, Transitions and Tempo were ranked lowest. The rankings seem to suggest that adjudicators placed greater emphasis on elements of music that allowed for musical interpretation from the performer rather than technical aspect of instrumental performance that are specified by explicit notation in the sheet music. Table 22 below provides a summary of the above findings.

Table 22: Summary of Findings

RESEARCH QUESTIONS	VARIABLES	DATA SOURCES	METHODS OF ANALYSIS	RESULTS
What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind certified FBA adjudicators?	Independent: *Adjudicator Group* Dependent: *Total Score*	FBA Concert Band MPA Performances Online Participant Surveys	Mann-Whitney Test Chi-Square Test of Independence Effect Size	Face-to-Face Mean Rank=3.5 Blind Certified Mean Rank=11.5 Z=-3.357 p<.01 φ=1.0 *r*=.59
What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-certified FBA adjudicators?	Independent: *Adjudicator Group* Dependent: *Total Score*	FBA Concert Band MPA Performances Online Participant Surveys	Mann-Whitney Test Chi-Square Test of Independence Effect Size	Face-to-Face Mean Rank=4.5 Blind Certified Mean Rank=11.2 Z=-3.107 p<.01 φ=.88 *r*=.55
What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-local certified adjudicators?	Independent: *Adjudicator Group* Dependent: *Total Score*	FBA Concert Band MPA Performances Online Participant Surveys	Mann-Whitney Test Chi-Square Test of Independence Effect Size	Face-to-Face Mean Rank=3.5 Blind Certified Mean Rank=11.5 Z=-3.357 p<.01 φ=1.0 *r*=.59
How do adjudicators rank the importance of the three major sub-captions and the criteria within each sub-caption?	Independent: *Adjudicators* Dependent: *Sub-Caption Rakings*	Online Participant Survey	Friedman Test Wilcoxon Test	Performance Fundamentals Median=1.0 μ=1.43,sd=.568 Technical Preparation Median=2.0 μ=2.40, sd=.621 Musical Effect Median=2.5 μ=2.17, sd=.913

A continued discussion of the results, conclusions, implications, delimitations and recommendations for future research will be presented in Chapter Five.

CHAPTER 5: DISCUSSION

Introduction

This study investigated adjudicator reliability and possible perceptual distortion in scores assessed by adjudicators at the Florida Bandmasters Association annual Music Performance Assessments (MPA). It investigated how adjudicators under conflicting sets of circumstances interpreted the criteria and rated musical performances. A sample of five concert band audio recordings from the FBA resource library were chosen and a sample of participants were selected to score the recordings using the criteria currently in use by the Florida Bandmasters Association. These participants were chosen from certified FBA concert band adjudicators, FBA members who are not certified concert band adjudicators and out of state judges who are certified though other judges association. Differences between groups were examined. In addition, data were collected on the participants' ranking of the musical criteria from the FBA concert band assessment instrument.

Statement of the Problem

To date, insufficient information exists concerning possible perceptual distortion in scores assessed by adjudicators at the annual Music Performance Assessments (MPA) which school music programs must attend in order to remain members of the Florida Bandmasters Association (FBA). Previous research has shown that factors such as director experience, stage presence and choice of repertoire can affect the outcome of a music

performance assessment rating. Bias also has been shown in situations where the adjudicator is familiar with the performer(s) or repertoire being performed (Bradley, 1972). In addition, Elliot (1995/1996) concluded that gender stereotypes associated with certain instruments also influenced an evaluator's perception of musical performance in smaller solo or chamber music settings.

Summary

A sample of five audio recordings from each of the top five Florida public school concert band directors who have the highest frequency of superior ratings at the Florida Bandmasters Association Music Performance assessment were collected. A sample of 10 concert-band certified Florida Bandmaster Association music adjudicators was selected. Ten FBA members, who are not concert band certified FBA adjudicators from the state of Florida, as well as 10 certified adjudicators from outside of Florida were also selected. A website link was sent to each of the 30 participants which included an MP3 recording of five separate state level, superior rated, concert band performances for their review using the Florida Bandmasters Association Concert Band MPA assessment instrument. An online survey corresponding to each of the five recordings, which contain the evaluation criteria to be used, was also provided to the adjudicators. Lastly, a final survey that asked the adjudicator to rank the sub-captions and the criteria within each sub-caption that are found on the Florida Bandmasters Association Concert Band MPA assessment instrument was provided. Information such as final ratings, and other musical element rankings were

collected and compared to the information and ratings given by the judges at the initial performance.

The Kruskal-Wallis Test found a statistically significant difference in the total score between the groups of adjudicators (χ^2 =20.97, *df*=3, *p*<.01). The mean ranks in order from greatest to least were Certified Non-Local (26.4), Certified FBA (23.3), Non-Certified FBA (14.5) and Face-to-Face Adjudicators (4.0) showing that there was a significant difference between the largest and smallest mean ranks. These three independent groups of blind adjudicators were each tested against the face-to-face adjudicators in an attempt to uncover where, and to what degree, differences in scoring existed.

A summary of findings has been offered around the four research questions that guided this study, and they are presented and discussed as they relate to the research and literature examined as part of this analysis.

Research Question 1

1. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind certified FBA adjudicators?

A Mann-Whitney post-hoc test was performed to determine if there was a statistical difference in total score between face-to-face and blind certified FBA adjudicators at a significance level of *p*=.01. The test results showed significant differences between the

groups (z=-3.357, p<.01). In addition, there was a significant difference in the mean rank of face-to-face adjudicators (3.5) and certified FBA adjudicators (11.5). A pairwise comparison was conducted and found a strong association between group and total score with Cramer's *V* reported as 1.0. Further, effect size value calculated at r=.59 suggested a strong practical significance between group and total score.

For the purpose of this study, a total score of five (5) from any individual adjudicator would be considered the best possible score, with that judge assessing a rating of one (1) to each of the five musical performances. Conversely, the worst total score that could possibly be given by any one adjudicator is twenty-five (25) with each performance being given a rating of five (5). In order for a recording to be contained in the FBA recording resource library, and considered for use in this study, it must have received a perfect score from any adjudicator who evaluated it at either the district, or state level during a face-to-face performance and assessment.

While all of the judges in this portion of the study were of the same larger population of certified FBA concert band adjudicators, the data clearly show that there was absolutely no agreement between the face-to-face and the blind adjudicator groups with respect to total score. FBA certified adjudicators who evaluated the live performance in no instance agreed with FBA certified adjudicators who were only presented with the audio recordings. While the face-to-face assessments gave the total score of five (5) in all cases, blind assessments ranged from a total score of seven (7), a relatively close to perfect score, all the way up to fourteen (14) in two cases, which were some of the highest total scores in the entire study. Lastly, Cramer's *V (1.0)* shows a complete association between group and

total score, in this case supporting the assumption that scores assessed by adjudicators are connected to the group and delivery method of the performance being evaluated.

Research Question 2

2. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-certified FBA adjudicators?

The results of this test showed significant differences between the groups (z=-3.107, p<.01). In addition, there was a significant difference in the mean rank of face-to-face adjudicators (4.0) and non-certified FBA adjudicators (11.2). A pairwise comparison was conducted and found a strong association between group and total score with Cramer's V reported as .88. Further, effect size value calculated at r=.55 suggested a strong practical significance between group and total score.

The data here show similar findings as previously reported in these independent groups of adjudicators. In this portion of the study, the original face-to-face adjudicator ratings were paired against blind non-certified FBA adjudicators. This sample was made up of members of the Florida Bandmasters Association, such as music educators, professionals and judges, which are not certified by FBA to evaluate at a concert band Music Performance Assessment. In one case, a member of the blind panel of judges gave a perfect score of five (5), however that is the only time this occurs in the entire study. Scores for this group were

generally better; with one blind non-certified FBA adjudicator assessing a score of six (6) and half of the blind non-certified FBA adjudicators give a total score of seven (7). Nonetheless one adjudicator again assessed a score of fourteen (14). Again, while not complete this time, a very strong association between total score and the adjudicator group was supported by Cramer's *V* at .88. From the data it is possible to conclude that the face-to-face and blind adjudicators' perceptions of the same musical performance were quite different.

Research Question 3

3. What is the difference, if any, between the scoring of a Florida Bandmasters Association Concert Band Performance by face-to face certified FBA adjudicators and blind non-local certified adjudicators?

Lastly, a final Mann-Whitney post-hoc test was performed to determine if there was a statistical difference in total score between face-to-face and certified non-local adjudicators at a significance level of p=.01. Again, the test results showed significant differences between the groups (z=-3.357, p<.01). In addition, there was a significant difference in the mean rank of face-to-face adjudicators (3.5) and certified non-local adjudicators (11.5). A pairwise comparison was conducted and found a strong association between group and total score with Cramer's *V* reported as 1.0. Further, effect size value calculated at r=.59 suggested a strong practical significance between group and total score.

Statistically identical to the blind certified FBA adjudicators, the blind non-local adjudicators also greatly differed from the face-to-face adjudicators in total scores assessed. Non-local certified adjudicators who evaluated the audio recording in no instance agreed with face-to-face FBA certified adjudicators who evaluated the live performance. As previously noted, the face-to-face assessments presented a total score of five (5) in all cases, however blind assessments from non-local certified adjudicators ranged from a total score of seven (7), all the way up to fifteen (15) in one case, which was the worst score in the entire study. Six adjudicators from this sample all gave total scores of twelve (12) or higher, as total scores skewed highest of any independent blind group. Finally, Cramer's *V (1.0)* shows another complete association between group and total score, supporting the conclusion that the face-to-face and blind non-local certified adjudicators' perceptions of the same musical performance are again quite different. From analysis of the data, it is reasonable to conclude that there is a strong difference in opinion on musical performances when presented as recorded examples as opposed to live performances.

Research Question 4

4. How do adjudicators rank the importance of the three major sub-captions and the criteria within each sub-caption?

A Freidman test was conducted to test for differences in medians among adjudicators for the three sub-captions contained on the Florida Bandmasters Association

concert band assessment sheet including Performance Fundamentals (median=1.0) Technical Preparation (median=2.0) and Musical Effect (median=2.5). The test was significant (χ^2 =15.27, *df*=2, *p*<.01) and a Kendall's coefficient of concordance of .25 suggested a medium difference in ranking among the three sub-captions. A follow up pairwise comparison was made through a Wilcoxon test at the *p*=.01 significance level. The concern for Performance Fundamental (μ=1.43, *sd*=.568) was greater than that of Technical Preparation (μ=2.40, *sd*=.621, *p*<.01) as well as Musical Effect (μ=2.17, *sd*=.913, *p*<.01). The concern for Technical Preparation did not differ significantly from Musical Effect (*p*>.01). This shows that adjudicators were more concerned with, and gave more weight to, the presence of fundamental training within the musical ensemble rather than the technical and musical precision of the actual performance being evaluated. In this instance an adjudicator is going to forgive some musical mistakes in a performance if it is obvious the ensemble is well trained in the basics of making good music, aligning with FBA philosophy (Florida Bandmasters Association Adjudication Manual, 2015).

The set of criteria contained in the Performance Fundamentals sub-caption include: Tone, Intonation, Balance, Blend, Sonority and Articulation. Tone was given the highest ranking by adjudicators (μ=1.63, *sd*=1.07) and ranked highest overall in 19 cases (*n*=30). Ranked lowest of the criteria was Articulation (μ=5.27, *sd*=1.11) placing lowest of the six criteria in 19 cases, 63.3% of the time. Again, the judges have favored tone quality, which is one of the most fundamental aspects of musical performance on an instrument, over other criteria. However, the problem here may be, as Burnsed, Hinkle & King (1985) found, that judges disagreed significantly in certain captions, with tone quality being the most notable.

Articulation is less of a performance decision of the individual performer as it is of the composer, and therefore mostly already notated in the music. This could lead to the reduced importance within this sub-caption placed on it by adjudicators.

The set of criteria contained in the Technical Preparation sub-caption include: Note Accuracy, Rhythmic Accuracy, Precision, Entrances, Releases, Interpretation, Clarity, Technique, Pulse, Dynamics and Transitions. Note Accuracy was given the highest ranking by adjudicators (μ=1.73, *sd*=1.34) and ranked highest overall in 18 cases (*n*=30). Ranked lowest of the criteria was Transitions (μ=9.73, *sd*=1.72) placing lowest of the eleven criteria in 13 cases, 43.3% of the time. This would seem logical, as performing the correct notes is one of the first technical aspects a musician learns, even going so far as to studying notation and pitch away from their instrument. Transitions represent the vaguest of the criteria in this sub-caption and could have led to its low rank by music adjudicators. Often there are very few, if any, musical transitions to speak of in a piece of concert literature. In addition, timing and performance of transitions is usually at the discretion of the conductor, rather than the performer.

The set of criteria contained in the Musical Expression sub-caption include: Expression, Shaping, Style, Interpretation, Phrasing, Tempo and Dynamics. Style was given the highest ranking by adjudicators (μ=3.03, *sd*=1.77) and ranked highest overall in 7 cases (*n*=30). It might be argued that performing in the correct style will guide the decisions made with respect to the other six criteria in the sub caption, hence the importance placed on it by adjudicators. Ranked lowest of the criteria was Tempo (μ=5.60, *sd*=1.94) placing lowest of the seven criteria in 14 cases, 46.7% of the time. The composer

usually places tempo markings in the music, and therefore there would be little opportunity for musical interpretation, if any, by the ensemble or director.

Typically, music adjudicators rate performances based on their own personal idea of quality and the importance of each musical element. Previous studies have shown that specific criteria used on a music evaluation instruments have not proven to be reliable (Burnsed, Hinkle & King, 1985). It is difficult to know to what degree a judge's score and opinion about how the ensemble performed coincides with the sub-caption criteria, as there is little room for feedback with respect to these performance standards. Jones (1986) and Winter (1993) went so far as to developed judges' sheets that include a Likert scale to gauge an adjudicator's level of agreement towards particular performance criteria. While the judging process involves human perceptions of musical characteristics, which can leave much room for interpretation, the sub-captions and sub-caption criteria contained on the FBA assessment instrument are vague and narrow in focus. More detailed measurement tools may be needed as music is complex and requires an equally complex measurement tool. Fiske (1975) suggested a successful musical performance is a united, cohesive phenomenon and detailed feedback is needed to properly assess it.

Conclusions

The information contained in this study is intended to provide information that could lead to development of a fair and balanced evaluation system for Florida Bandmasters Association Music Performance Assessments. Based on the review of related

literature and the data collected during this study, the following conclusions were reached by the researcher:

1. There is a statistically significant difference in scores assessed by face-to-face adjudicators versus blind adjudicators, possibly attributed to the halo effect. While face-to-face adjudicators all agreed on straight superior ratings for the performances included in this study, none of the blind participants, excluding one, gave the same ratings. This indicates that some mitigating factor or piece of biographical datum that was not present in the blind audio recordings may have caused a discrepancy in the scores assessed by the two groups. Halo effect, as described by Feeley (2002) is an evaluator's tendency to overemphasize the relationship between a subject's traits or behaviors and may have been a factor in the original face-to-face assessments. The recordings used were those of the music directors who had the highest frequency of state level superior rated music programs, and quite possibly better reputations in the Florida Bandmasters Association community. Face-to-face adjudicators would of course know exactly which music programs (and directors) they were adjudicating during a live performance, while blind adjudicators did not have any of this information. Here, the inclination of the halo effect might allow the director's reputation to have a positive influence on total scores assessed by face-to-face adjudicators (Blum and Naylor, 1968).
2. Some other qualitative aspects of the live performance are being observed and are creating a perceptual distortion during the musical evaluation. Studies have shown

that factors that are non-musical in nature can come into consideration when evaluation a musical performance. In a series of research studies VanWeelden (2002) found that female directors with a thin build were higher rated in musical performance than those with a larger build. Another of his studies concluded that concert bands performing African American Spirituals conducted by African American conductors were rated higher than ensembles led by white conductors even though the musical performances provided to the judges were identical. In addition, judges rated white conductors higher with respect to the western concert band literature, leading VanWeelden to the conclusion that the judges may have racially stereotyped the conductors. Elliott (1995/1996) discovered that typical gender stereotyping of instruments often influenced a judge's perception of a musical performance. In the case of these studies, the musical performance given to the judges was the same throughout, heightening the fact that gender was a consideration in the musical evaluations. Morrison et al. (2009) discovered one might judge an ensemble's musicality based on the expressiveness of the onstage director. In another example Davis (2000) studied the size of bands and found a positive correlation between the size of the band and the rating it achieved at a music evaluation festival, with larger bands receiving higher ratings. Vines, Krumhansl, Wanderley, and Levitin, (2006) found a relationship between the visual movement of the performer and the music phrasing that is perceived, while Juchniewicz (2008) establish that evaluators gave higher ratings on musical criteria such as dynamics, rubato and phrasing to those performers who incorporated full

movement of their body to their performance. It was also reported by Thompson, Graham, and Russo (2005) that performers could communicate expressiveness through facial expressions, in turn enhancing an evaluator's listening experience.

3. Adjudicator training and professional organization membership may lead to more consistent music performance assessment results. In this study, the assumption of normality for the dependent variable total score was met for both Certified FBA Adjudicators and Certified Non-Local Adjudicators, and statistically the results they produced were the same (z=-3.357, p<.01). Mean rank of face-to-face certified adjudicators (3.5) and other blind certified adjudicators (11.5) were identical as well. This statistical similarity was not found in the sample of non-certified adjudicators. Qualified adjudicators work under certain constraint and are trained to use specific methods when assessing a performance using an assessment instrument's sub-captions and criteria. The assumption at the FBA state-level concert band MPA might be that the performing bands are all at the top end of the spectrum, but blind adjudicators, not knowing what performance they are listening too, might simply fall back on their training and score the performance more as they see fit, not prescribing to typical contest dynamics or norms. Bradley (1972) found that a factor such as a judge's training, experience, and knowledge of repertoire all strongly affect the outcome of the performance assessment. The Florida Bandmasters Association has created guidelines in which an FBA member can become a certificated judge. As outlined in the FBA handbook, after having seven years of teaching experience and after receiving straight superior ratings three out

of the last five years, a music director may apply to become certified (FBA Adjudication Handbook 2014-2015, 2014). At this point, the internship process begins, where candidates attend official training and shadow other certified judges during a number of FBA sponsored Music Performance Assessments, comparing their assessments and ratings with those of the certified judges on the panel. At the culmination of this year long process the candidate's materials are sent to the FBA Executive Board for approval, and they will be added to the list of official judges used by FBA for events (FBA Adjudication Handbook 2014-2015, 2014). While not within the scope of this studies research, other professional music judges associations around the county have similar application and training requirements.

4. The current Florida Bandmasters Association Music Performance Assessment adjudication sheets are too qualitative in nature to be used for formal teacher evaluations. Additional research would need to be done to develop a better system. It is not outrageous to assert that musical performance might suffer as teachers become more focused with standardized test preparation when salary, benefits and job security are at stake. According to Pistone, (2012) there are problems with the current models of assessment for hard-to-measure subjects, such as music. While a standardized test might be able to measure a student's knowledge and understanding about the fundamental concepts of music, it in no way can measure a student's ability to perform or compose music. Pistone continues on to say that such a standardized test will also not be able to show if a teacher has success in educating students in performing music as an ensemble. Music teachers in the state of Florida

have suggested that performance events judged by an independent panel of adjudicators are the most appropriate way to test music student achievement, even going so far as to naming such current music festivals Music Performance Assessments (Cochran-Smith, 2007). Music directors place a large importance on these music festivals, and in the age of accountability, a director's future may be based on the outcome. Critics have argued however that these music assessments offer no baseline pretest and cannot track individual student achievement or individual student learning (Fiske 1983).

The above conclusions serve as another step in a body of knowledge that investigates the ways in which music performance assessments and festivals can become a more valid and reliable method of assessing a director's success as an educator. In turn, in the age of increased accountability, these data can be used towards a broader purpose such as teacher assessments and VAM scores for educators in more non-traditional classroom settings, which are traditionally more difficult for an administrator to assess.

Delimitations and Recommendations for Future Study

Data collected and literature reviewed during this study point towards the following recommendations for future analysis:

1. Continuing research into other aspects of the Florida Bandmasters Association Music Performance Assessment, such as solo and ensemble festival, jazz and

marching MPA, and the reliability of those scores assessed. While, recordings of jazz band and marching band performances are usually made, there is no current library on file as there is for state superior rated concert band performances. Solo and ensemble festival performances are not typically recorded, so in order to conduct such a study recordings and original ratings assessed to students would have to be documented by the researcher in order to evaluate them against the opinion of a blind research study group.

2. Studying the effect of the literature selected by the director on scores assessed by adjudicators. Although concert ensembles are required to choose the music they perform for evaluation from a FBA approved list, some music on that list is considered (and marked) "significant literature", and adjudicators might possibly evaluate those pieces differently when rating a music program or performance.
3. Continued investigation on how biographical data of the director such as age, gender, race, name, level of education or years of experience might result in any possible perceptual distortion, relationship (either positive or negative) or halo effect at a Florida Bandmasters Association Music Performance Assessment. In addition, investigating how information such a music program's school name or reputation might also result in any possible perceptual distortions or significant differences in scores assessed.
4. Examining pairwise comparisons between the multiple blind adjudicator groups, rather than with face-to-face adjudicators. While the data showed significant differences between the face-to-face adjudicators and the three independent blind

groups, additional analysis might show no such statistical difference between blind groups tested.

5. An investigation into what specific information about a director or music program triggers any perceptual distortions or leads to a difference in total score assessed by an adjudicator. In this case, a single recorded performance might be evaluated by several independent groups of adjudicators, each receiving a separate and distinct piece of information. An attempt would be made to isolate what knowledge might lead to perceptual distortions, halo effect or statistically higher scores.
6. A comparison of the scores assessed in the different sub-captions by individual judges against the musical criteria they rated as most important. This may serve to alert adjudicators to their own personal biases when scoring a musical performance.
7. Comparing the results of one individual adjudicator and musical performance against varying versions of concert band adjudication sheets to test if a broader scale, stricter definition of the sub-captions or change in wording of criteria would result in a significant difference in score assessed.

APPENDIX A: IRB APPROVAL LETTER

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Raymond A. Donato**

Date: **April 08, 2015**

Dear Researcher:

On 4/8/2015, the IRB approved the following human participant research until 04/07/2016 inclusive:

Type of Review:	UCF Initial Review Submission Form
Project Title:	A STUDY ON THE INFLUENCE OF PERCEPTUAL DISTORTION IN THE SCORING OF MUSICAL PERFORMANCES BY FLORIDA BANDMASTERS ASSOCIATION ADJUDICATORS
Investigator:	Raymond A Donato
IRB Number:	SBE-15-11115
Funding Agency:	
Grant Title:	
Research ID:	n/a

The scientific merit of the research was considered during the IRB review. The Continuing Review Application must be submitted 30days prior to the expiration date for studies that were previously expedited, and 60 days prior to the expiration date for research that was previously reviewed at a convened meeting. Do not make changes to the study (i.e., protocol, methodology, consent form, personnel, site, etc.) before obtaining IRB approval. A Modification Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at https://iris.research.ucf.edu .

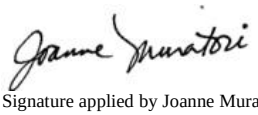
If continuing review approval is not granted before the expiration date of 04/07/2016, approval of this research expires on that date. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

Use of the approved, stamped consent document(s) is required. The new form supersedes all previous versions, which are now invalid for further use. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Participants or their representatives must receive a copy of the consent form(s).

All data, including signed consent forms if applicable, must be retained and secured per protocol for a minimum of five years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained and secured per protocol. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual.

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:



Signature applied by Joanne Muratori on 04/08/2015 01:27:53 PM EDT

IRB Coordinator

APPENDIX B: IRB CONSENT FORM

A Study On The Influence Of Perceptual Distortion In The Scoring Of Musical Performances By Florida Bandmasters Association Adjudicators

Informed Consent

Principal Investigator: Raymond A. Donato

Faculty Advisor: Kenneth Murray, J.D., Ph.D.

Introduction:

Researchers at the University of Central Florida (UCF) study many topics. To do this we need the help of people who agree to take part in a research study. You are being invited to take part in a research study which will include about 30 people nationally. You have been asked to take part in this research study because you are either a music educator or certified music adjudicator. You must be 18 years of age or older to be included in the research study.

The person doing this research is Raymond A. Donato, a graduate student from the University of Central Florida department of Teaching, Learning, and Leadership. Because the researcher is a graduate student, he is being guided by Dr. Kenneth Murray, a UCF faculty advisor in the department of Teaching, Learning, and Leadership.

What you should know about a research study:

- Someone will explain this research study to you.
- A research study is something you volunteer for.
- Whether or not you take part is up to you.

- You should take part in this study only because you want to.
- You can choose not to take part in the research study.
- You can agree to take part now and later change your mind.
- Whatever you decide it will not be held against you.
- Feel free to ask all the questions you want before you decide.

Purpose of the research study:

The purpose of this study is to test the reliability of the Music Performance Assessment ratings sheets used by the Florida Bandmasters Association during a concert band festival. There is a need to examine the criteria contained on these assessment instruments, and how an adjudicator under a contrasting set of circumstances interprets them, which might affect the outcome of a concert band's final assigned rating.

There are a few research publications that focus on some observable elements such as the race of the director, ensemble uniform choice, the directors conducting style or even the stage presence of the musicians and how these components can affect the perception of an ensemble's musical performance. However, there is little to no research that simply tests the validity of the Music Performance Assessment Ratings Sheets used by the Florida Bandmasters Association during a concert band festival. This study may provide valuable information that could lead to better development of a fair and balanced rating system.

What you will be asked to do in the study:

When participating in this study, you will be asked to listen to a musical performance, and evaluate the performance using an online questioner. From your own personal computer, you will be directed to a website that will guide you through the process. On this website you will

find five musical examples, and five corresponding links to answer questions about the musical examples. You do not have to complete all the surveys in one sitting, you will have 30 days to complete all of them. Be sure to respond to each survey only once.

1. Click on the musical excerpt to listen to it directly from the provided webpage.
2. Click on the corresponding link that will take you to an online survey where you will answer questions about the performance you just heard. As the survey will open in a separate browser window, you will be able to listen to the musical performance and respond to the survey simultaneously, as you might be doing during a typical concert band evaluation. Please be sure to press "submit" at the end of each survey before closing the window.
3. Repeat the process for the remaining musical excepts and corresponding surveys (a total of five).
4. Click on the last survey link to answer some general questions about your professional musical beliefs.

Location:

The research will take place through a webpage where the participants can listen to the musical performances to be evaluated and fill out an online evaluation form. The music excerpts and online forms can be accessed from any public or private computer with an internet connection. This can be done at the participants leisure in any setting.

Time required:

We expect that you will be in this research study for approximately two (2) hours. This time can

be divided into multiple sessions as desired by the participant.

Risks:

Potential risks to you may include:

- Loss of time – (time to complete the evaluation will take approximately two hours).
- Mental Fatigue.
- Frustration.

Benefits:

There are no expected benefits to you for taking part in this study.

Compensation or payment:

There is no compensation or other payment to you for taking part in this study.

Confidentiality:

We will limit your personal data collected in this study to people who have a need to review this information. We cannot promise complete secrecy.

Study contact for questions about the study or to report a problem: If you have questions, concerns, or complaints, or think the research has hurt you, please contact:

Raymond A. Donato
Graduate Student, University of Central Florida department of Teaching, Learning, and Leadership
(561) 414-3786
rdonato@knights.ucf.edu

or

Dr. Kenneth Murray
Faculty Supervisor, University of Central Florida department of Teaching, Learning, and Leadership
(407) 823-1468
murray@mail.ucf.edu

IRB contact about your rights in the study or to report a complaint:

Research at the University of Central Florida involving human participants is carried out under the oversight of the Institutional Review Board (UCF IRB). This research has been reviewed and approved by the IRB. For information about the rights of people who take part in research, please contact: Institutional Review Board, University of Central Florida, Office of Research & Commercialization, 12201 Research Parkway, Suite 501, Orlando, FL 32826-3246 or by telephone at (407) 823-2901. You may also talk to them for any of the following:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You want to get information or provide input about this research.

Withdrawing from the study:

If you decide to leave the study, contact the investigator so that the investigator can remove your incomplete data from the study and select a new participant in your place.

The person in charge of the research study can remove you from the research study without your approval. If a randomly selected participant is found to have any prior attachment to or affiliation with the music excerpts being evaluated, they may be withdrawn from the study. The participant will be notified in writing that they have been removed from the study. We will tell you about any new information that may affect your health, welfare or choice to stay in the research.

APPENDIX C: MUSICAL EXAMPLE EVALUATION FORM

Music Example #1 Evaluation Form

* Required

Please enter your Participant ID number. *

You should have received this number from the researcher along with your link to this survey.

Please rate the overall PERFORMANCE FUNDAMENTALS of this musical performance on a scale of 1 (highest quality) to 5 (lowest quality). *

Considering the following criteria: Tone Quality, Intonation, Balance, Blend, Band Sonority and Physical Articulation.

- 1 - Superior
- 2 - Excellent
- 3 - Good
- 4 - Fair
- 5 - Poor

Please rate the overall TECHNICAL PREPARATION of this musical performance on a scale of 1 (highest quality) to 5 (lowest quality). *

Consider the following criteria: Note accuracy, Rhythmic Accuracy, Precision, Entrances, Releases, Interpretive Articulation, Clarity of Articulation, Technique, Stability of Pulse, Dynamics Observed and Transitions.

- 1 - Superior
- 2 - Excellent
- 3 - Good
- 4 - Fair
- 5 - Poor

Please rate the overall MUSICAL EFFECT of this musical performance on a scale of 1 (highest quality) to 5 (lowest quality). *

Consider the following criteria: Expression, Shaping of Line, Style, Interpretation, Phrasing, Tempo and Dynamic Expression.

- 1 - Superior
- 2 - Excellent
- 3 - Good
- 4 - Fair
- 5 - Poor

Please add any ADDITIONAL PERFORMANCE COMMENTS below, if applicable:

Submit

Never submit passwords through Google Forms.

APPENDIX D: MUSICAL ELEMENTS ORDER OF IMPORTANCE SURVEY

Musical Elements Order of Importance Survey

* Required

Please enter your Participant ID number. *

You should have received this number from the researcher along with your link to this survey.

When evaluating a concert band performance, rate these musical criteria in order of importance from 1 (most important) to 3 (least important). *

	1 - Most Important	2	3 - Least Important
Performance Fundamentals			
Technical Preparation			
Musical Effect			

When evaluating a concert band performance, rate these PERFORMANCE FUNDAMENTALS criteria subdivisions in order of importance from 1 (most important) to 6 (least important). *

	1 - Most Important	2	3	4	5	6 - Least Important
Tone Quality						
Intonation						
Balance						
Blend						
Band Sonority						
Physical Articulation						

When evaluating a concert band performance, rate these TECHNICAL PREPERATION criteria subdivisions in order of importance from 1 (most important) to 11 (least important). *

	1 - Most Important	2	3	4	5	6	7	8	9	10	11 - Least Important
Note accuracy											
Rhythmic Accuracy											
Precision											
Entrances											
Releases											
Interpretive Articulation											
Clarity of Articulation											
Technique											
Stability of Pulse											
Dynamics Observed											
Transitions											

When evaluating a concert band performance, rate these MUSICAL EFFECT criteria subdivisions in order of importance from 1 (most important) to 7 (least important). *

	1 - Most Important	2	3	4	5	6	7 - Least Important
Expression	○	○	○	○	○	○	○
Shaping of Line	○	○	○	○	○	○	○
Style	○	○	○	○	○	○	○
Interpretation	○	○	○	○	○	○	○
Phrasing	○	○	○	○	○	○	○
Tempo	○	○	○	○	○	○	○
Dynamic Expression	○	○	○	○	○	○	○

Submit

Never submit passwords through Google Forms.

APPENDIX E: FBA ASSESSMENT INSTRUMENT

Florida Bandmasters Association

Adjudicator's Comment Sheet

CONCERT BAND

School:_____

Classification:_____ Performance Time:_____ Date:_____

Selections: 1._____
2._____
3._____

PERFORMANCE FUNDAMENTALS				TECHNICAL PREPARATION				MUSICAL EFFECT			
	1	2	3		1	2	3		1	2	3
Tone Quality	___	___	___	Note accuracy	___	___	___	Expression	___	___	___
Intonation	___	___	___	Rhythmic Accuracy	___	___	___	Shaping of Line	___	___	___
Balance	___	___	___	Precision	___	___	___	Style	___	___	___
Blend	___	___	___	Entrances	___	___	___	Interpretation	___	___	___
Band Sonority	___	___	___	Releases	___	___	___	Phrasing	___	___	___
Physical Articulation	___	___	___	Interpretive Articulation	___	___	___	Tempo	___	___	___
				Clarity of Articulation	___	___	___	Dynamic Expression	___	___	___
				Technique	___	___	___				
				Stability of Pulse	___	___	___				
				Dynamics Observed	___	___	___				
				Transitions	___	___	___				
	☐	☐	☐		☐	☐	☐		☐	☐	☐
FINAL:	☐			FINAL:	☐			FINAL:	☐		

Officials will include a + or – by the subdivisions, which mean they are noticeably good or noticeably needing improvement as related to the letter grade assigned. The absence of any marks indicates a performance consistent with the letter assigned. After completing the previous, enter an A, B, C, D, or E to indicate the level of performance in each category. Average the three letter grades in each category to arrive at a final letter grade. Average the three final grades to arrive at the FINAL RATING.

COMMENTS

(Including: Stage Presence, Discipline, Posture, Instrumentation, Strong Points, Weak Points – Continue on Reverse Side-→)

Recommended For: _____
(Superior, Excellent, Good, Fair, Poor)
Write out Final Rating

Adjudicator's Signature

Rev 12/10

LIST OF REFERENCES

A History of FBA. (2014). Retrieved June 26, 2014 from http://flmusiced.org/fba/dnn/About-FBA/History

Abeles, H. F. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education*, 21, 246-255.

Abeles, H. F. (2010). The historical contexts of music education. In H. Abeles & L. Custodero, Critical issues in music education: Contemporary theory and practice (9). New York: Oxford Press.

About FSMA. (2014). Retrieved June 25, 2014 from http://fsma.flmusiced.org/about/

Abril, C. R & Gault, B. M. (2006). The state of music in the elementary school: The principal's perspective. *Journal of Research in Music Education*, 54(1), 6-20.

Abril, C. R. & Gault, B. M. (2008). The state of music in secondary schools: The principal's perspective. *Journal of Research in Music Education*, 56(1). 68-81.

Aguilar, C. A. (2011). *The development and application of a conceptual model for the analysis of policy recommendations for music education in the united states in the Department of Music Education of the Jacobs School of Music* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (UMI Number: 3456436).

Alicke, M. D., & Govorun, O. (2005). The better than average effect. In M. D. Alicke, D. A. Dunning, & J. I. Krueger (Eds.), *The self in social judgment* (pp. 85–106). New York: Psychology Press.

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal and Social Psychology,* 41, 258-290.

Asmus, E. P. 1999. Music assessment concepts. *Music Educators Journal* 86 (2): 19.

Assumptions-of-the-Factorial-Anova. Retrieved February 2, 2015 from http://www.statisticssolutions.com/assumptions-of-the-factorial-anova/

Austin, J. R. (1988). The effect of music contest format on self-concept, motivation, achievement, and attitude of elementary band students. *Journal of Research in Music Education*, 36(2), 95-107.

Azzara, C. D. (1993). The effect of audiation-based improvisation techniques on the music achievement of elementary music students. *Journal of Research in Music Education*, 41, 328-342.

Banks, W. P., & Krajicek, D. (1991). Perception. *Annual Review of Psychology,* 305.

Barresi, A. L. & Olson, G. (1992). The nature of policy and music education. In R. Colwell (Ed.), *Handbook of Research on Music Teaching and Learning* (pp. 760-772). New York: Schirmer Books.

Bassin, W. M. (1974). A note on the biases in students' evaluations of instructors. *Journal of Experimental Education*, 43, 16–17.

Bauer, W. I. (1993). The relationship between rehearsal procedures and contest ratings for high school bands. *Contributions to Music Education*, 20, 32-44.

Bazan, D. E. (2007). *Teaching and learning strategies used by student-directed teachers of middle school band* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (A, 65(05) UMI No. 3264513).

Beaver, M. E. (1973). *An investigation of personality and value characteristics of successful high school band directors in North Carolina* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (A, 34(05) UMI No. 7326392).

Bergee, M. J. (1993). A comparison of faculty, peer, and self-evaluation of applied brass jury performances. *Journal of Research in Music Education*, 41, 19-27.

Bergee, M. J., & Platt, M. C. (2003). Influence of selected variables on solo and small-ensemble festival ratings. *Journal of Research in Music Education*, 51 (4), 342-353.

Berry, B., Hoke, M., & Hirsch, E. (2004) The search for highly qualified teachers. *The Phi Delta Kappan*, 85, no. 9: 685.

Blum, M. I., and Naylor, J. C. (1968). *Industrial psychology: Its theoretical and social foundations.* New York: Harper & Row.

Bordia, R., & DiFonzo, N. (2005). Psychological motivations in rumorspread. In G. A. Fine, C. Heath, & V. Campion-Vincent (Eds.), *Rumor mills: The social impact of rumor and legend* (pp. 87–101). New Brunswick, NJ: Transaction.

Bradley, I. L. (1972). Effect on student musical preference of a listening program in contemporary art music. *Journal of Research in Music Education,* 20(3), 344-353.

Brakel, T. D. (1997). *Attrition of instrumental music students as a function of teaching style and selected demographic variables* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (A, 58(12), 4592. UMI No. 9816946).

Bräm, P. B., & Braem, T. (2001). A pilot study of the expressive gestures used by classical orchestra conductors. *Journal of the Conductor's Guild*, 22(1–2), 14–29.

Burnsed, V., Hinkle, D., & King, S. (1985). Performance evaluation reliability at selected concert band festivals. *Journal of Band Research*, 21(1), 22-29.

Burnsed, V., & King, S. (1987). How reliable is your festival rating? *Update: Applications of Research in Music Education*, 5 (3), 12-13.

Burnsed, V., Sochinski, J., & Hinkle, D. (1983). The attitude of college band students toward high school marching band competition. *Journal of Band Research*, 19(1), 11-17.

Buss, D. M. (1987). Selection, evocation, and manipulation. *Journal of Personality and Social Psychology*, 53(6), 1214–1221.

Byo, J. (1990). Recognition of intensity contrasts in the gestures of beginning conductors. *Journal of Research in Music Education*, 38, 157–163.

Byo, S. J. (1999). Classroom teachers' and music specialists' perceived ability to implement the national standards for music education. *Journal of Research in Music Education*, 47(2), 111-123.

Caimi, F. J. (1981). Relationships between motivation variables and selected criterion measures of high school band directing success. Journal of Research in Music Education, 29(3), 183-198.

Carla, A. E. (2011). The development and application of a conceptual model for the analysis of policy recommendations for music education in the united states in the Department of Music Education of the Jacobs School of Music. Indiana University

Chaiken, S. (1987). The heuristic model of persuasion. In M. P. Zanna, J. M. Olson, & C. P. Herman (Eds.), *Social influence: The Ontario Symposium* (Vol. 5, pp. 3–39). Hillsdale, NJ: Erlbaum.

Clark, A. (1997). *Being there.* Cambridge, MA: MIT Press.

Clark, A. E., & Kashima, Y. (2007). Stereotypes help people connect with others in the community: A situated functional analysis of the stereotype consistency bias in communication. *Journal of Personality and Social Psychology*, 93(6), 1028–1039.

Clarke, E. F. (2005). *Ways of listening: An ecological approach to the perception of musical meaning*. New York: Oxford University Press.

Cochran-Smith, M. (2007). Teacher education: Where are we and where are we going? In M. Schmidt, Collaborative Action for Change (15). New York: Rowan & Littlefield Education.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: a metaanalysis of multi-section validity studies. *Review of Educational Research,* 51 (3), 281–309.

Colwell, R. (1994). Aggressive educational policy and MENC. *The Quarterly Journal of Music Teaching and Learning*, 5(2), 50-62.

Colwell, R. (1999). The 1997 assessment in music: Red flags in the sunset. *Arts Education Policy Review,* 100 (6): 33–39.

Colwell, R. (2005). Whither programs and arts policy? *Arts Education Policy Review*, 106(6), 19-29.

Colwell, R. (2006). Music teacher education in this century. *Arts Education Policy Review*, 108(1), 15-27.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, 90, 218–244.

Corder, G. & Foreman, D. (2009). *Nonparametric statistics for non-statisticians*. Hoboken: John Wiley & Sons. pp. 99–105.

Costello, J. D. (2005). *Classroom management in music ensembles: Exploring the relationship between perceived classroom management skills and performance achievement* (Masters thesis). Retrieved from: Masters abstracts international. (44(01), 61. UMI No. 1431270).

Cozby, P. C. *(2009). Methods in Behavioral Research 10th ed.* Boston: McGraw-Hill Higher Education.

Craik, K. H. (2008). *Reputation: A network interpretation*. New York: Oxford University Press.

Croft, J. (1984). Current problems & concerns of band directors. *The School Musician*, 55(8), 36-37.

Crone, D. T. (2002). *A historical descriptive analysis of federal, state, and local education policy and its influence on the music education curriculum in the New York City Public Schools, 1950 – 1999* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (UMI No. ATT 3041878).

Darby, J. A. (2007). Are course evaluations subject to a halo effect? *Research in Education, 77*, 46-55.

Davidson, J. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music*, 21, 103–113.

Davis, A. P. (1998). Performance achievement and analysis of teaching during choral rehearsals. *Journal of Research in Music Education*, 46(4), 496-509.

Davis, R. B. (2000). *A Study of the relationship between rehearsal procedures and contest ratings for high school marching band* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (A, 61(03), 925. UMI No. 9965728).

Dawes, B. L. (1989). *A survey of Alabama band directors regarding marching band competitions and music performance achievement* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (A, 51(04), 1037. UMI No. 9025298).

Denrell, J. (2005). Why most people disapprove of me: Experience sampling in impression formation. *Psychological Review*, 112, 951–978.

Diehl, D. (2007). *Factors related to the integration of the national standards in the secondary school wind band* (Doctoral Dissertation). Retrieved from Dissertation abstracts international. (UMI No. AAT 3255053).

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics* 6 (3): 241–252.

Edgar S. (2012). Communication of expectations between principals and entry-year instrumental music teachers: Implications for music teacher assessment. *Arts Education Policy Review*, 113 (2012): 137.

Eisner, E. (1985). *Learning and Teaching the Ways of Knowing, 84th Yearbook of the National Society for the Study of Education, Part II.* Chicago: University of Chicago Press.

Elliott, C. A. (1995/1996). Race and gender as factors in judgments of musical performance. *Bulletin of the Council for Research in Music Education,* 127, 50-56.

Fallis, T. L. (1999). Standards-based instruction in rehearsal. *Music Educators Journal,* 85(4), 18-23.

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: Valence asymmetries. *Journal of Personality and Social Psychology,* 87(3), 293–311.

FBA Handbook 2014-2015. (2014). Retrieved June 26, 2014 from http://flmusiced.org/fba/dnn/About-FBA/FBA-Handbook-Constitution-Bylaws-Information

Feeley, T. (2002). Comment on Halo Effects in Rating and Evaluation Research. *Human Communication Research*, *28*(4), 578-86.

Feldman, J. M. (1986). A note on the statistical correction of halo error. *Journal of Applied Psychology,* 71, 173–176.

Fiedler, K. (2000). Beware of samples! A cognitive–ecological sampling approach to judgment biases. *Psychological Review*, 107(4), 659–676.

Fiese, R. K. (1991). The relationship among conductor's rankings of three unfamiliar wind bandscores. *Journal of Research in Music Education*, 39 (3), 239-247.

Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement*, 14, 419–429.

Fisicaro, S. A., & Vance, R. J. (1994). Comments on the measurement of halo. *Educational and Psychological Measurement*, 54, 366–371.

Fiske, H. E. (1975). Judge-group differences in the rating of secondary school trumpet performances. *Journal of Research in Music Education*, 23, 186-196.

Fiske, H. E. (1977). Who's to judge: New insights into performance adjudication. *Music Educators Journal,* 64, 23-25.

Fiske, H. E. (1983). Judging musical performances: Method or madness? Update: *The Applications of Research in Music Education*, 1 (3), 7-10.

Florida Bandmaster Association Adjudication Committee Report. (2011). Retrieved January 13, 2012 from http://flmusiced.org/fba/dnn/Portals/0/Bandmaster/May%20Board%20Meeting%202011%20(2).pdf

Florida Bandmasters Association Adjudication Manual. (2015). Retrieved January 23, 2016 from http://fba.flmusiced.org/media/1274/adjudication-manual-2015.pdf

Florida Bandmaster Association District Meeting #2 Minutes. (2012). Retrieved June 21, 2014 from http://www.flmusiced.org/fba/dnn/Portals /0/MeetingMinutes%5C2012-2013_District12_Meeting2.pdf

Florida Bandmaster Association District Meeting #4 Minutes. (2013). Retrieved June 21, 2014 from https://flmusiced.org/fba/dnn/Portals/0/MeetingMinutes%5C2012-2013_District9_Meeting4.pdf

Florida School Music Association Bylaws. (2014). Retrieved June 25, 2014 from http://fsma.flmusiced.org/media/1108/fsma-bylaws-amended-october-2011.pdf

Fonder, M. & Eckrich, D.W. (1999). A survey on the impact of the voluntary national standards on American college and university music teacher education curricula. *Bulletin of the Council for Research in Music Education*, 140, 28-40.

Fosse, J. B. (1965). *The prediction of teaching effectiveness: An investigation of the relationship among high school band contest ratings, teacher characteristics, and school environment factors* (Doctoral dissertation). Retrieved from: Dissertation

abstracts international. (A, 26(06), 3391. UMI No. 6512081)

Foster, E. K. (2004). Research on gossip: Taxonomy, methods, and future directions. *Review of General Psychology*, 8(2), 78–99.

Fredrickson, W. E., Johnson, C. M., & Robinson, C. R. (1998). The effect of preconducting and conducting behaviors on the evaluation of conductor competence. *Journal of Band Research*, 33 (2), 1-13.

Frequently Asked Questions About FSMA. (2014). Retrieved June 25, 2014 from http://fsma.flmusiced.org/about/frequently-asked-questions/

Gardner, H. (1983). *Frames of Mind.* New York: Basic Books.

Garnett, L. (2005). Research report: Gesture, style and communication. *Master Singer*, 55, 14–15.

Geringer, J. M., Cassidy, J. W., & Byo, J. L. (1997). Nonmusic majors' cognitive and affective responses to performance and programmatic music videos. *Journal of Research in Music Education*, 45, 221–233.

Gibson, E. J. (1988). Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge (english). *Annu. Rev. Psychol., 39*, 1-41.

Gilbert, D. T. (1998). Ordinary personology. In D. T. Gilbert & S. T. Fiske (Eds.), *The handbook of social psychology* (4th ed., Vol. 2, pp. 89–150). New York: McGraw-Hill.

Gilbert, D. T., & Jones, E. E. (1986). Perceiver-induced constraint: Interpretations of self-generated reality. *Journal of Personality and Social Psychology*, 50(2), 269–280.

Gillespie, R. (1998). National standards for successful school string and orchestra teachers. *American String Teacher*, 48(3), 30-31.

Goertz, M., & Duffy, M. (2003). Mapping the landscape of high-stakes testing and accountability programs. *Theory into Practice,* 42 no. 1: 4.

Goodstein, R. E. (1984). *An investigation into leadership behaviors and descriptive characteristics of band directors in the United States* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (A, 45(08), 2433. UMI No. 8424644).

Goodstein, R. E. (1987). An investigation into leadership behaviors and descriptive characteristics of band directors in the United States. *Journal of Research in Music Education*, 35(1), 13-25.

Gordon, M. E. (1980). Organizational behavior: A managerial and organizational perspective (book). *Personnel Psychology, 33*(1), 228-231.

Grechesky, R. N. (1985). *An analysis of nonverbal and verbal conducting behaviors and their relationship to expressive music performance* (Doctoral dissertation). Retrieved from: Dissertation Abstracts Internationa. (48, 2656A.)

Greene, Maxine. (1988). *The Dialectic of Freedom*. New York: Teachers College Press.

Groulx, T. J. (2009). *Are band ratings more closely associated with the band director or the school?* Manuscript submitted for publication.

Groulx, T. J. (2010). *An examination of the influence of band director teaching style and personality on ratings at concert and marching band events* (Doctoral dissertation). Retrieved from: Dissertation abstracts International. A, 71(11). (UMI No. 3425686).

Guegold, W. K. (1989). *An analysis of the adjudication results in the 1986-1988 Ohio Music Education Association State Marching Band Finals with an emphasis on adjudicator*

consistency (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (A, 50(09), 2821. UMI No. 9006128).

Gumm, A. J. (2003a). *Music teaching style*. Galesville, MD: Meredith Music Publications.

Gumm, A. J. (2003b). The effects of choral music teacher experience and background on music teaching style. *Visions of Research in Music Education*, 3(February), 6-22.

Gumm, A. J. (2007). Using a generic student opinion survey to evaluate college conductors: Investigation of validity, dimensionality, and variability. *Bulletin of the Council for Research in Music Education*, 171(4), 37-50.

Hale, C. D., Herreid, C., & Waugh, G. (1996). Assessing teaching effectiveness in a liberal arts college: The student perspective.

Hamann, D. L. (1990). Classroom environment as related to contest ratings among high school performing ensembles. *Journal of Research in Music Education,* 38(3), 215-224.

Hanna, W. (2007). The new Bloom's taxonomy: Implications for music education. *Arts Education Policy Review,* 108(4), 7-16.

Harris, B. P. (1991). *Comparisons of attained ratings to instructional behaviors and techniques exhibited by band directors in sight-reading performance situations* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (A, 52(08), 2852. UMI No. 9202299)

Harrison, S. D., Lebler, D., Carey, G., Hitchcock, M., & O'Bryan, J. (2013). Making music or gaining grades? Assessment practices in tertiary music ensembles. *British Journal of Music Education, 30(1), 27-42.*

Heimonen, M. (2006). Justifying the right to music education. *Philosophy of Music Education Review*, 14(2), 119-141.

Heine, S. J. (2001). Self as cultural product: An examination of East Asian and North American selves. *Journal of Personality*, 69(6), 881–906.

Henninger, J. C. (2008). The effects of performance quality ratings on perceptions of instrumental music lessons. *Update: Applications of Research in Music Education, 27*(1), 9-16.

Higgins, E. T., & Rholes, W. S. (1978). Saying is believing: Effects of message modification on memory and liking for the person described. *Journal of Experimental Social Psychology*, 14, 363–378.

Hinckley, J. (1997). Implementing the national k-12 music standards. *Proceedings: Association of Schools of Music*, 86, 77-81.

Hoffa, H. (1988). Arts education and politics: The odd coupling, *Design for Arts in Education*, 89(5), 2-12.

Hoffman, M.E. (1994). MENC: Policy, advocacy, and enlightened self-interest. *The Quarterly Journal of Music Teaching and Learning*, 5(2), 44-49.

Hogwood, B. W. & Gunn, L. A. (1984). *Policy analysis for the real world*. Oxford: Oxford University Press.

Holden, R. B. (2010). *Face validity*. In I. B. Weiner & W. E. Craighead, *The Corsini Encyclopedia of Psychology (4th ed.)* (pp. 637-638) Hoboken, New Jersey: Wiley.

Hope, S. (1989). National Conditions and Policy Imperatives, *Design for Arts in Education*, 91(1), 15-35.

Hope, S. (2002). Policy frameworks, research and K-12 schooling. In R. Colwell & C. Richardson (Eds.), *The New Handbook of Research on Music Teaching and Learning* (pp. 5-16) New York: Oxford University Press.

Hope, S. (2007). Strategic policy issues and music teacher preparation. *Arts Education Policy Review*, 109(1), 3-10.

Hourigan, R. (2011). Race to the top: Implications for professional development in arts education. *Arts Education Policy Review*, 112, 60.

House, R. E. (2000, March). *Effects of expressive and nonexpressive conducting on advanced instrumentalists*. Paper presented at the National MENC In-Service Conference, Washington, DC.

Howes, M. B. (1990). *The psychology of human cognition : Mainstream and genevan traditions.* New York, NY: Pergamon Press, 1990; 1st ed.

Juchniewicz, J. (2008). The influence of physical movement on the perception of musical performance. *Psychology of Music*, 36, 417–427.

Kelley, H. H. (1950). The warm-cold variable in first impressions of persons. *Journal of Personality*, 18, 431-439.

Kenny, D. A., Albright, L., Malloy, T. E., & Kashy, D. A. (1994). Consensus in interpersonal perception: Acquaintance and the Big Five. *Psychological Bulletin,* 116(2), 245–258.

Kenny, D. A., & Kashy, D. A. (1994). Enhanced co-orientation in the perception of friends: A social relations analysis. *Journal of Personality and Social Psychology*, 67(6), 1024–1033.

Kenny, D. A., Mohr, C. D. & Levesque, M. J. (2001). A social relations variance partitioning of dyadic behavior. *Psychological Bulletin*, 127(1), 128–141

Kerchner, J. L. (2001). Incorporating the national standards in performing classes. *Teaching Music,* 9(1).

Kerr, S. P. (2002). *Legal responsibilities and rights of music educators: An investigation of cases, court verdicts, and legislation* (Doctoral Dissertation). Retrieved from: Dissertation abstracts international. (UMI No. ATT 3060357).

Kirkland, N. J. (1996). *South Carolina schools and Goals 2000: National standards in Music* (Doctoral Dissertation). Retrieved from Dissertation abstracts international. (UMI No. ATT 9623096).

Kivy, Peter. (1991). Music and the Liberal Education. *Journal of Aesthetic Education 25(3), 79-93.*

Kos, R. P. (2007). Incidental change: The influence of educational policy implementation on music education programs and practice. (Doctoral dissertation). Retrieved from: Dissertation abstracts international.

Kozlowski, S. W. J., Kirsch, M. P., & Chao, G. T. (1986). Job knowledge, rate familiarity, conceptual similarity, and halo error: An exploration. *Journal of Applied Psychology*, 71, 45–49.

Kruskal-Wallis-Test. Retrieved February 2, 2015 from http://www.statisticssolutions.com/kruskal-wallis-test/

Laib, J. R. (1993). *The effect of expressive conducting on band performance* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (54, 3258A.)

Lambourne, P. E. (2002). *Classroom policy and educational practices: A study of primary classroom music education in Kern County, California*. (Doctoral Dissertation). Retrieved from: Dissertation abstracts international. (UMI No. ATT 3074944).

Langer, Susanne K. (1982). *Feeling and Form: A Theory of Art Developed from Philosophy in a New Key.* Baltimore: Johns Hopkins Press.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.

LeCroy, H. F. (1998). Community-based music education: Influences of industrial bands in the American south. *Journal of Research in Music Education*, 46(2), 248-64.

Lee, P. T. (1997). Implementing the national k-12 music standards: *California. Proceedings: National Association of Schools of Music*, 86, 86-88.

Leimer, M. (2012). *Female band directors and adjudicators in Florida* (Doctoral dissertation). Electronic theses, treatises and dissertations. Paper 4980.

Linn, R. L. 2003. Accountability: Responsibility and reasonable expectations. *Educational Researcher* 32 (7): 3–13.

Louk, D. P. (2002). *National Standards for music education: General music teachers' attitudes and practices* (Doctoral Dissertation). Retrieved from Dissertation abstracts international. (UMI No. ATT 3042585).

Lowry, R. (2015). *Subchapter 15a.The Friedman Test for 3 or More Correlated Samples.* Retrieved on January 27, 2016 from http://vassarstats.net/textbook/ch15a.html

Lucas, K. V.,Hamann, D. L., & Teachout, D. J. (1996). Effect of perceptual mode on the identification of expressiveness in conducting. *Southeast Journal of Music Education*, 8, 166-175.

Lund, A. (2013). *Association – Part III.* Retried on January 29, 2016 from https://statistics.laerd.com/premium/sts/sts-association-3.php

Malloy, T. E., Albright, L., Kenny, D. A., Agatstein, F., & Winquist, L. (1997). Interpersonal perception and metaperception in nonoverlapping social groups. *Journal of Personality and Social Psychology*, 72(2), 390–398.

Mantie, R. (2012). Striking up the band: Music education through a foucaultian lens. *Action, Criticism, and Theory for Music Education*, 11(1), 99-123.

McCook, W. M. (1976). *A multivariate study of variables effecting student ratings of teaching and course outcomes within a multiple instructor mode.* Paper presented at the annual meeting of the National Council of Measurement in Education. San Francisco, Ca. April 19-23.

McCrae, R. R., & Costa, P. T. (2003). *Personality in Adulthood (2nd Ed.)*. New York: Guilford Press.

McCurry, M. L. (1998). *Handchime performance as a means of meeting selected standards in the national standards of music education* (Doctoral Dissertation). Retrieved from: Dissertation abstracts international. (UMI No. ATT 9828359).

McIntyre, R. A. (1990). *Legal issues in the administration of public school music Programs* (Doctoral Dissertation). Retrieved from Dissertation abstracts international. (UMI No. ATT 9119097).

McLaughlin, M. W. (2006). Implementation research in education: Lessons learned, lingering questions, and new opportunities. In M. I. Honig (Ed.) *New Directionsin Education Policy Implementation: Confronting Complexity*. Albany, NY: State University of New York Press.

McPherson, G. E., & Thompson, W. F. (1998). Assessing music performance: issues and influences. *Research Studies in Music Education*, 10(1), 12-24.

Meyer, L. B. (1956). *Emotion and Meaning in Music*. Chicago: Chicago University Press.

Mi-Young, O., & Jyotika, R. (2003). Halo-effect: conceptual definition and empirical exploration with regard to South Korean subsidiaries of US and Japanese multinational corporations. *Journal of Communication Management,* 7 (4), 317–30.

Mishook, J. J., & M. L. Kornhaber. 2006. Arts integration in an era of accountability. *Arts Education Policy Review* 107 (4): 3–11

Morrison, S. J., Price, H. E., Geiger, C. G., & Cornacchio, R. A. (2009). The effect of conductor expressivity on ensemble performance evaluation. *Journal of Research in Music Education, 57*(1), 37-49

Murphy, K. R., & Anhalt, R. L. (1992). Is halo error a property of the rater, ratees, or the specific behaviors observered? *Journal of Applied Psychology*, 77, 494–500.

Murphy, K. R., & Reynolds, D. H. (1988). Does true halo affect observed halo? *Journal of Applied Psychology,* 73, 1–4.

Mursell, James L. (1934). *Human Values in Music Education*. New York: Silver Burdett.

Music Educators National Conference. *(1991). Growing up complete: The imperative for music education. The report of the national commission on music education.*

Myers, D. E. (2002). Policy issues in connecting music education with arts education. *The new handbook of research on music teaching and learning*, 909–30. New York: Oxford University Press.

Nash, M. A. (2013). Cultivating our "musical bumps" while fighting the "progress of popery": The rise of art and music education in the mid-nineteenth century united states. *Educational Studies: Journal of the American Educational Studies Association*, 49(3), 193-212.

Neilson, J. (1973). A blueprint for adjudicators. *The Instrumentalist*, 28(5), 46-48.

Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: Outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, 53(3), 431–444.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

O'Brien, M. L. (1992). Using Rasch procedures to understand psychometric structure in measures of personality. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 61-76). Norwood, NJ: Ablex.

Orman, E. K. (2002). Comparisons of the national standards for music education and the music specialists' use of class time. *Journal of Research in Music Education,* 50(2), 155-164.

Overview of Florida's Teacher Evaluation System. (2016). Retrieved January 12, 2016 from http://www.fldoe.org/core/fileparse.php/7503/urlt/0102688-overviewfloridasteacherevaluationsystem.pdf

Owen, S. A., & Connecticut Univ, Storrs Bureau of Educational Research, and Service. (1976). *The validity of student ratings: A critique.*

Papageorgi, I., Creech, A., Haddon, E., Morton, F., De Bezenac, C., Himonides, E., & Welch, G. (2010). Perceptions and predictions of expertise in advanced musical learners. *Psychology of Music, 38*(1), 31-66.

Performance Evaluation. (2016). Retrieved January 12, 2016 from http://www.fldoe.org/teaching/performance-evaluation

Perrine, W. M. (2013). Music Teacher Assessment and Race to the Top: An Initiative in Florida. *Music Educators Journal*, *100*(1), 39-44.

Phelps, L., & Others, A. (1986). The effects of halo and leniency on cooperating teacher reports using likert-type rating scales. *Journal of Educational Research, 79*(3), 151-54.

Philosophy and Purpose of the FBA. (2014). Retrieved June 26, 2014 from http://flmusiced.org/fba/dnn/About-FBA/Philosophy

Phoenix, Philip H. (1964). *Realms of Meaning: A Philosophy of the Curriculum for General Education.* New York: McGraw-Hill Book Co.

Pike, G. R. (1999). The constant error of the halo in educational outcomes research. *Research in Higher Education,* 40, 61–86.

Pistone, N. *Envisioning Arts Assessment*. Retrieved January 27, 2013 from https://cfaefl.org/AssessmentProject/userfiles/Envisioning%5FArts%5FAssessment.pdf.

Pohlmann, J. T. (1975). A multivariate analysis of selected class characteristics and student

ratings of instructions. *Multivariate Behavioral Research,* 10 (1), 81–91.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher,* 31 (7): 3–14.

Price, H. E. (1983). The effect of conductor academic task presentation, conductor renforcement, and ensemble practice on performers' musical achievement, attentiveness, and attitude. *Journal of Research in Music Education*, 31(4), 245-257.

Price, H. E. (2004). Mapping music education research in the USA: A response to the UK. *Psychology of Music*, 32(3), 322-329.

Price, H. E., & Chang, E. C. (2001). Conductor expressivity and ensemble performance: An exploratory investigation. *Contributions to Music Education*, 28(2), 9–20.

Price, H. E., & Chang, E. C. (2005). Conductor and ensemble performance expressivity and state festival ratings. *Journal of Research in Music Education, 53*(1), 66-77.

Price, H. E., & Winter, S. (1991). Effect of strict and expressive conducting on performances and opinions of eighth grade students. *Journal of Band Research*, 27(1), 30–43.

Purohit, A., Magoon, A. J., & Delaware Univ., N. (1971). *The validity of student-run course evaluations.* Retrieved from https://login.ezproxy.net.ucf.edu/login?auth=shibb&url=http://search.ebscohost.com.ezproxy.net.ucf.edu/login.aspx?direct=true&db=eric&AN=ED047630&site=ehost-live

Race to the Top Assessments. Florida Department of Education. Retrieved January 25, 2012 from http://www.fldoe.org/arra/racetothetop/assessments/

Race to the Top for Student Success Act – SB 736. (2011). Retrieved June 21, 2014 from http://feaweb.org/_data/files/2011/PPA/736/What_you_need_to_know_about_SB_736.pdf

Rae, L. (1997). *How to Measure Training Effectiveness (third edition).* Aldershot: Gower Publishing.

Rae, L. (2002). *Assessing the Value of your Training: the Evaluation Process from Training Needs to the Report to the Board*. Aldershot: Gower Publishing.

Read, Herbert. (1958). *Education Through Art*. London: Faber and Faber.

Rebell, M., & Hunter, M. (2004). Highly qualified teachers: Pretense or legal requirement? *The Phi Delta Kappan*, 85, no. 9: 691.

Reis, H. T., Nezlek, J., & Wheeler, L. (1980). Physical attractiveness in social interaction. *Journal of Personality and Social Psychology*, 38(4), 604–617.

Reis, H. T., Senchak, M., & Solomon, B. (1985). Sex differences in the intimacy of social interaction. *Journal of Personality and Social Psychology*, 48(5), 1204–1217.

Review and Approval Checklist for RTTT Teacher Evaluation Systems. Retrieved January 27, 2013 from http://www.fldoe.org/finance/contracts-grants-procurement/american-recovery-reinvestment-act/teacher-principal-evaluation-sys.stml

Richmonds, J. W. (1992). Arts education as equal educational opportunity: The legal issues. *Journal of Research in Music Education*, 40(3), 236-252.

Rickels, D. A. (2008). A comparison of variables in Arizona marching band festival results. *Journal of Band Research*, 44(1), 25-39.

Riveire, J.H. (1997*). California string teachers' curricular content and attitudes regarding*

improvising and the national standards (Doctoral Dissertation). Retrieved from: Dissertation abstracts international. (UMI No. ATT 9835075)

Robbins, P., & Aydede, M. (Eds.). (2008). *Cambridge handbook of situated cognition*. New York: Cambridge University Press.

Ryan,C., & Costa-Giomi, E. (2004) ;Attractiveness bias in evaluation of young pianists' performances. *Journal of Research in Music Education*, 52, 141-154.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, 88, 413–428.

Saul, C. E. (1976). *An analysis of the relationship of selected characteristics of Mississippi public high school band directors, students, and programs to their festival ratings* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (A, 37(12), 7602. UMI No.7711761).

Saunders, T. C., & Holahan, J. M. (1997). Criteria-specific rating scales in evaluation of high school instrumental performance. *Journal of Research in Music Education* , 45 (2), 259-272.

Scheib, J. W. (2006). Policy implications for teacher retention: Meeting the needs of the dual identities of arts educators. *Arts Education Policy Review*, 107(6), 5-10.

Schnieder, D. J. (. (1991). Social cognition (english). *Annu. Rev. Psychol., 42*, 527-561.

Scholl, J. C. (2005). Helping students explore three styles of learning to illustrate the perception process. *Communication Teacher, 19*(2), 53-56.

Schopp, S. E. (2006). *A study of the effects of national standards for music education, number 3, improvisation, and number 4, composition on high school band instruction in New*

York State (Doctoral Dissertation). Retrieved from: Dissertation abstracts international. (UMI No. ATT 3225193).

Schwadron, Abraham A. (1967). *Aesthetics: Dimensions for Music Education.* Washington, D.C.: Music Educators National Conference.

Scott, S. J. (2012). Rethinking the roles of assessment in music education. *Music Educators Journal, 98*(3), 31-35.

Scott, T. B. (1996). *The construction of a holistic, criterion-referenced sight-singing test for high school sopranos based on the voluntary national standards for music education* (Doctoral Dissertation). Retrieved from: Dissertation abstracts international. (UMI No. ATT 9712431).

Shaum M., & Ganson, H. (2005). The No Child Left Behind Act of 2001: The federal government's role in strengthening accountability for student performance. *Review of Research in Education*, 29: 157.

Sheldon, D. A. (1994). The effects of competitive versus noncompetitive performance goals on music students' ratings of band performances. *Bulletin for the Council of Research in Music Education,* 121, 29-41.

Sheldon, D. A. (2000). Effects of music expression and conductor disposition on school musicians' affect. *Quadreni della SIEM: Seinestrale di Ricma e Didattica Musical,* 16, 287-293.

Shove, P., & Repp, B. (1995). *Musical motion and performance: Theoretical and empirical perspectives.* In J. Rink (ed.) The practice of performance. Cambridge University Press, 1995.

Shuler, S.C. (2001). Music and education in the twenty-first century: A retrospective. *Arts Education Policy Review*, 102(3), 25-36.

Sidoti, V. J. (1990). *The effects of expressive and nonexpressive conducting on the performance accuracy of selected expression markings by individual high school instrumentalists* (Doctoral dissertation). Ohio State University.

Siegel, C. (1988). *Nonparametric statistics for the behavioral sciences* (Second ed.). New York: McGraw–Hill.

Silvey, B. A. (2009). The effects of band labels on evaluators' judgments of musical performance. *Update: Applications of Research in Music Education, 28*(1), 47-52.

Silvey, B. A. (2011). The effect of ensemble performance quality on the evaluation of conducting expressivity. *Journal of Research in Music Education, 59*(2), 162-173.

Sims, H. P., & Lorenzi, P. (1992). *The new leadership paradigm : Social learning and cognition in organizations.* Newbury Park, CA: Sage Publications.

Skube, J. T. (2002). *Implementation of the national standards for music education within secondary instrumental music programs in the state of Michigan* (Masters Thesis). Retrieved from: Masters abstracts international. (UMI No. ATT 1411982).

Small, C. (1998). Musicking: *The meanings of performing and listening. Hanover* NH: Wesleyan University Press.

Smith, E. R., & Collins, E. C. (2009). Contextualizing person perception: Distributed social cognition. *Psychological Review, 116*(2), 343-364.

Smith, J. W. (1999). *Correlation of discrete and continuous contest ratings with marching band director rehearsal behaviors* (Doctoral dissertation). Retrieved from:

Dissertation abstracts international. (A, 60(09), 3303. UMI No. 9946123).

Smith, P. K., & Trope, Y. (2006). You focus on the forest when you're in charge of the trees: Power priming and abstract information processing. *Journal of Personality and Social Psychology*, 90(4), 578–596.

Smith, R. A. (1987). *Discipline-based Art Education: Origins, Meaning, and Development.* Urbana and Chicago: University of Illinois Press.

Snyder, D. (2001). The national standards in junior high band rehearsals. *Teaching Music*, 8(6).

Snyder, M., Tanke, E. D., & Berscheid, E. (1977). Social perception and interpersonal behavior: On the self-fulfilling nature of social stereotypes. *Journal of Personality and Social Psychology*, 35(9), 656–666.

Spurlock, H. L. (2002). *The impact of student-centered pedagogy and students' feelings of autonomy, competence, and relatedness on motivation: Implications for test motivation and test performance* (Doctoral dissertation). Retrieved from: Dissertation abstracts international. (A, 63(01), 88. UMI No. 3040827).

Spurrier, J. D. (2003). On the null distribution of the Kruskal–Wallis statistic. *Journal of Nonparametric Statistics* 15 (6): 685–691.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48(6), 1467–1478.

Sullivan, T. M. (2003). *Factors influencing participation of Arizona high school marching bands in regional and state festivals* (Doctoral dissertation). Retrieved from:

Dissertation abstracts international. (A, 64(02), 388. UMI No. 3080892).

Tang, T. L., & Tang, T. L. (1987). *A correlational study of students' evaluations of faculty performance and their self-ratings in an instructional setting.*

Teachout, D. J. (1997). Preservice and experienced teachers' opinions of skills and behaviors important to successful music teaching. *Journal of Research in Music Education*, 45(1), 41-50.

Tellstrom, A. T. 1971. *Music in American education past and present.* New York: Holt, Reinhart, and Winston.

Thompson, W. F., Graham, P., & Russo, F. A. (2005). Seeing music performance: Visual influences on perception and experience. *Semiotica*, 156, 203–227.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology,* 4, 25–9.

Thorndike, E. L., & Hagen, E. (1977). *Measurement and Evaluation in Psychology and Education (second edition)*. New York: Wiley.

Thorne, A. (1987). The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality and Social Psychology*, 53(4), 718–726.

Van Weelden, K. (2002). Relationships between perceptions of conducting effectiveness and ensemble performance. *Journal of Research in Music Education*, 50, 165–176.

VanPatten, B. W. (1997). *A model curriculum for a high school instrumental music program implementing the "National Standards for Arts Education"* (Masters Thesis). Retrieved from: Masters abstracts international. (UMI No. ATT 1387659)

VanWeelden, K. (2004). Racially stereotyped music and conductor race: perceptions of performance. *Bulletin of the Council for Research in Music Education,* 160(2), 38-48.

Vanweelden, K., & McGee, I. R. (2007). The influence of music style and conductor race on perceptions of ensemble and conductor performance. *International Journal of Music Education, 25*(1), 7-17.

Wagner, M. J. (1991). The effect of adjudicating three videotaped popular music performances on a "composite critique" rating and an "overall" rating. *Missouri Journal of Research in Music Education*, 28, 53-70.

Walsh, K. (2004). Through the looking glass: How NCLB's promise requires facing some hard truths about teacher quality. *The Clearing House*, 78, no. 1: 22.

Wang, C. C. & Sogin, D. W. (1997). Self-reported versus observed classroom activities in elementary general music. *Journal of Research in Music Education*, 45(3), 444 - 456.

Wapnick, J., Darrow, A. A., Kovacs, J., & Dalrymple, L. (1997). Effects of attractiveness on evaluation of vocal performance. *Journal of Research in Music Education*, 45, 470-479.

Wapnick, J., Mazza, J. K., & Darrow, A. A. (1998). Effects of performer attractiveness, stage behavior, and dress on violin performance evaluation. *Journal of Research in Music Education*, 46, 510–521.

Wapnick, J., Mazza, J. K., & Darrow, A. A. (2000). Effects of performer attractiveness, stage behavior, and dress on evaluation of children's piano performances. *Journal of Research in Music Education*, 48, 323–336.

Washington, K. E. (2007). *A Study of Selected Characteristics of Mississippi High School Bands and Band Festival Ratings* (Doctoral dissertation). Retrieved from: Dissertation

abstracts international. (A, 68(08). UMI No. 3280886).

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications (Structural analysis in the social sciences).* New York: Cambridge University Press.

Wegener, D. T., & Petty, R. E. (1997). The flexible correction model: The role of naive theories of bias in bias correction. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 29, pp. 141–208). Mahwah, NJ: Erlbaum.

Wells, A. S. (1993). The sociology of school choice: Why some win and others lose in the educational marketplace. In E. Rassell & R. Rothstein (Eds.) *School Choice: Examining the Evidence*. Washington, D.C.: Economic Policy Institute.

Widmeyer, W. N., & Loy, J. W. (1988). When you're hot, you're hot! Warm-cold effects in first impressions of persons and teaching effectiveness. *Journal of Educational Psychology, 80*(1), 118-21.

*W*inerip, M. (2012, Januray 22). In Obama's race to the top, the dirty work is left to those on the bottom. *New York Times.*

Winter, N. (1993) Music performance assessment: A study of the effects of training and experience on the criteria used by music examiners. *International Journal of Music Education*, 22, 34-39.

Woodbridge, William C. (1831). *A Lecture on Vocal Music as a Branch of Common Education.* Delivered in the Representatives' Hall, Boston, August 24, 1830, before the American Institute of Instruction. Boston: Hilliard, Gray, Little and Wilkins.

Yarbrough, C., & Madsen, K. (1998). The evaluation of teaching in choral rehearsals. *Journal of Research in Music Education*, 46(4), 469-481.

Yatani, K. (2014). *Kruskal-Wallis and Friedman Tests.* Retrieved from January 29, 2016 from http://yatani.jp/teaching/doku.php?id=hcistats:kruskalwallis